7-14-2020

Automatic Depression Screening and Depressive Symptom Prediction Using Smartphone Sensing Data

Shweta Ware
University of Connecticut - Storrs, shweta.ware@uconn.edu

Automatic Depression Screening and Depressive Symptom Prediction Using Smartphone Sensing Data

Shweta Ware, Ph.D.

University of Connecticut, 2020

ABSTRACT

Depression is a common, yet serious health problem. It has significant detrimental impacts on both physical and psychological functioning. Current diagnosis techniques rely on physician-administered or patient self-administered interview tools, which are burdensome and suffer from recall bias. Additionally, these techniques incur higher medical costs. There is an urgent need for an accurate, objective and easily accessible depression screening tool for mass usage. In this dissertation, we explore the usage of smartphone sensing data, collected directly on smartphones or meta-data collected from a WiFi infrastructure, for automatic depression screening and depressive symptom prediction.

In the first part of the dissertation, we develop a novel approach that investigates the feasibility of automatic large-scale depression prediction using meta-data captured in an institution's WiFi network, without direct data capture (i.e., running apps) on phones. Specifically, when smartphones connect to a WiFi network, their locations (and hence the locations of the users) can be determined by the access points that they associate with; the location information over time provides important insights into

the behavior of the users, which can be used for depression screening. To investigate the feasibility of this approach, we have analyzed two datasets, each collected over several months, involving tens of participants recruited from a university. Our results demonstrate that WiFi meta-data is effective for passive depression screening: the F_1 scores are as high as 0.85 for predicting depression, comparable to those obtained by using sensing data collected directly from smartphones.

In the second part of the dissertation, we explore the feasibility of using smartphone sensing data for automatic prediction of all major categories of depressive symptoms, including both cognitive (in interests, mood, concentration) and behavioral (in appetite, energy level, sleep) symptoms. Specifically, we consider two types of smartphone data, one collected passively on smartphones and the other collected from an institution's WiFi infrastructure and construct a family of machine learning based models for the prediction. Both scenarios require no efforts from the users and can provide objective assessment of depressive symptoms. Our results demonstrate that smartphone data can be used to predict both behavioral and cognitive symptoms effectively, with F_1 score as high as 0.86. Our study makes a significant step forward over existing studies, which only focus on predicting the overall depression status (i.e., whether one is depressed or not).

In the third part of the dissertation, we explore the impact and influence of social interaction related features on mental health and wellness. Specifically, using a family of machine learning models, we have used Short Message Service (SMS) and call record data for depression prediction. By using the social interaction data collected via an Android phone app from college age students. Our results demonstrate that social interaction data can be used to predict depression effectively, with F_1 score as high as 0.80. Our results show that the people with depression spend more time on the

Shweta Ware, Ph.D.
University of Connecticut, 2020

phone calls than the non-depressed population. We also found that majority of the depressed individuals send and receive fewer number of messages and communicate with a limited number of contacts.

Automatic Depression Screening and Depressive Symptom Prediction Using Smartphone Sensing Data

Shweta Ware

M.S., University of Connecticut, 2015

B.Tech, National Institute of Technology, Raipur, 2009

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2020

APPROVAL PAGE

Doctor of Philosophy Dissertation

Automatic Depression Screening and Depressive Symptom Prediction Using Smartphone Sensing Data

Presented by

Shweta Ware, B.Tech, M.S.

Major Advisor _____
Dr. Bing Wang

Associate Advisor _____
Dr. Jinbo Bi

Associate Advisor _____
Dr. Alexander Russell

University of Connecticut

2020

ACKNOWLEDGMENTS

The past five years have been most significant in my life with constant learning at each step that gradually improved me as a student, a researcher and a person. From the bottom of my heart, I am really thankful to my advisor Dr. Bing Wang for her constant support and motivation over the past five years for my Ph.D. study and research. Words can't be enough to express how grateful I am for all the inspiration and motivation I have got from her that makes me what I am today. Her passion for research has taught me the importance of high-quality research and I got some wonderful opportunities to work and collaborate with other researchers. I value her time, patience and understanding throughout this journey of my Ph.D. study. It is really my honor to be her student.

I'd like to specially thank Dr. Jinbo Bi and Dr. Alexander Russell for their valuable guidance and comments on my research and dissertation. I feel fortunate and honored to have had the opportunity to work closely with them. A very special thanks also goes to Dr. Qian Yang and Dr. Suining He for their help and guidance on my dissertation. I'd also like to thank Dr. Jayesh Kamath who has been our research collaborator for the LifeRhythm project.

I am also grateful to all my friends who always supported and encouraged me. I'd like to thank my present and past fellow lab mates, Dr. Abdurrahman Arikan, Dr. Asma A. Farhan, Chaoqun Yue, Chinmaey Shende, Cheonjin Park, Levon Nazaryan,Dr. Ozgur Oksuz, Dr. Ruofan Jin, Stephen Sam, Reynaldo Morillo, Dr. Yuexin Mao, Yanyuan Qin, for all their help during my Ph.D. study. I am glad I

made some great friends during my time at UConn. I'd like also thank Dr. Jin Lu and Chao Shang for all their support while working on the research together.

I'd like to thank my family for all their love and support. I'd like to thank my parents for always making me believe that dreams do come true with hard work, patience and honesty. I'd like to especially thank my husband Aanay Ware for all his love, support and patience. He always believed in me and he has always walked by my side over all these past years. My Ph.D. study wouldn't have been possible without him. I am also thankful to my son Atharv Ware whose love has been my biggest strength all the time.

Contents

Chapter 1

Introduction and Background

1.1 Overview

Depression is a serious mental illness that has fatal impacts on both physical and mental health. It incurs higher medical costs and also results in higher mortality. The existing depression diagnosis methods are clinician administered or patient self administered. These methods not only require higher efforts and cost but also, often suffer from recall bias. Also, the lack of trained professionals (14.5 psychiatrists per 100,000) further result in worse consequences. In some countries, people still have the stereotype thinking where they don't want to talk openly about a sensitive topic like mental health. As a result, mental health problems remain undiagnosed and often lead to serious consequences. To effectively address depression as a public health problem, there is an urgent need for an accurate, objective and easily-accessible depression screening tool for mass usage.

1.1.1 Smartphone Data for Mental Health Applications

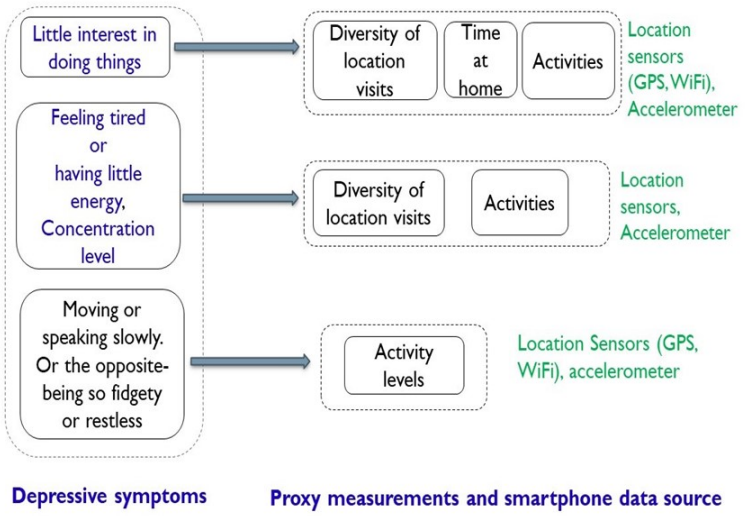


FIGURE 1.1: Overview of depressive symptoms and proxy measurements using smartphone data.

With the emergence of mobile computing technologies, smartphone sensing data applications are proliferating in several new directions. One of the main contributions of ubiquitous computing using smartphone data is its application for tracking mental health and wellness. There are numerous reasons like ubiquitous adoption of smartphones, their rich sets of embedded sensors like GPS, WiFi, SMS, phone call records, accelerometer, gyroscope, screen etc., that make them an ideal platform for recording and assessing the behavioral patterns of individuals. Smartphones thus serve as "human sensor" for capturing and analyzing human behavior.

Figure 1.1 illustrates an overview of how various behavioral and cognitive depressive symptoms can be inferred using several proxy measurements that can be calculated using smartphone data. For example, a person's interest levels can be

assessed in terms of the diversity of places he/she visits over a period of time, time spent at home and their activity levels. Similarly, if a person is feeling tired or has little energy and concentration levels, then that person may visit less places and will be less active. Another depressive symptom that assesses if someone is moving or speaking slowly than usual or vice versa, then that could also be gauged by the person's activity levels. We observe that there is a relationship between these depressive symptoms and day-to-day behavioral patterns in terms of location visits and activities. These behavioral traits can be recorded and measured by using smartphone sensors like GPS, WiFi, accelerometer etc. To summarize, behavioral proxy measurements recorded and calculated using smartphone sensing data can be used for predicting the depression.

Recent studies have used smartphone sensing data for depression prediction. Existing studies require running an app in the background that passively collects smartphone sensing data (using an array of sensors like GPS, WiFi, accelerometer, phone usage patterns via SMS, phone call records, email history etc.). After recording the sensing data, several useful behavioral features are calculated. These features are then fed to pre-trained machine learning models for depression prediction. These pre-trained models are constructed during the training phase using the ground truth data i.e. depression status from clinical assessments or user self-reports (like PHQ9 and QIDS). In what follows, we describe various high-level approaches used by the existing studies for depression and stress prediction using smartphone data.

Using location visit information. Some studies have used location visit information as a primary source for assessing the depression and stress among individuals [31, 13, 25, 87, 49, 71, 81]. Canzian and Musolesi [13] studied the relationship between the mobility patterns and depression, and found that individualized ma-

chine learning models outperformed general models. Farhan et al. [25] found that the features extracted from the smartphone sensing data can predict depression with good accuracy. Yue et al. [87] investigated fusing GPS and WiFi association data, both collected locally on smartphones, for more complete location information for improved depression detection. Lu et al. [49] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. Suhara et al. [71] developed a deep learning based approach that forecasts severely depressive mood based on self-reported histories.

Using social interaction information. Social interaction plays a significant role in the day-to-day life of individuals. Social ties play a beneficial role in the maintenance of psychological well-being [41]. Several studies have investigated the usage of various social interaction data sources like SMS, call, email and webhistory for mood and depression prediction [58, 59, 26, 77, 12, 47, 72, 27]. Studies [64, 46] have demonstrated the role of interpersonal communication and social relationships on mental health of individuals. Razavi et al. [58] have found that participants with depression spend more time on their mobile devices to make and receive fewer and shorter calls, and send more text messages than participants without depression. They trained an array of machine learning classification algorithms for depression prediction. The best model was a random forest classifier, which had an out-of-sample balanced accuracy of 0.768. The balanced accuracy increased to 0.811 when participants' age and gender were included. Rohani et al. [59] conducted a systematic review to provide an overview of the correlations between objective behavioral features and depressive mood symptoms. They used 7 categories of features, of which, social interaction category included call duration and frequency, missed calls, number of incoming and outgoing text messages along with the character count of messages.

Faurholt-Jepsen et al. [27] found that the people with severe depressive symptoms receive more calls, makes less calls and answer less calls. Also, they found that in case of more manic symptoms, there are more outgoing text messages per day, phone calls per day are longer and characters in received messages are lower.

Using phone usage information. Saeb et al. [61] found significant correlation between the phone usage and mobility patterns with respect to the self-reported Patient Health Questionnaire-9 (PHQ-9) scores. Wang et al. [81] studied the impact of workload on stress and day-to-day activities of students. They found significant correlation between the behavioral traits (in terms of conversation duration, number of locations visited, sleep) and depressive mood. Wang et al. [82] proposed depressive symptom features (that are calculated from phone and wearable passive sensor data) that proxy 5 out of the 9 major depressive disorder symptoms defined in the diagnostic manual (DSM-5) for college students. They found that depressed individuals' phone usage is higher at study places, have irregular sleep schedules and also visit fewer places during the day. Xu et al. [85] presented a new method based on association rule mining for generating contextually filtered features using passive mobile and wearable data. They used a variety of smartphone data including bluetooth, call, screen usage, location, campus map, sleep and steps for feature extraction. They showed that the best rules selected by their method are highly interpretable and can capture students' routine behaviors, and behavior pattern differences with and without depressive symptoms. Most of the studies discussed above use location and activity for depression and depressive symptoms prediction.

1.2 Topics in this Dissertation

After discussing the role of smartphone data for mental health applications followed by the existing state-of-the-art, we now present the specific problems we address in this dissertation. We investigate the following three problems for automatic depression screening using smartphone data: 1) explore an approach that is available at a low cost to a large scale population for automatic depression screening; 2) using smartphone data for providing a detailed picture on depression conditions; 3) exploring the role of social interaction behavior for depression prediction.

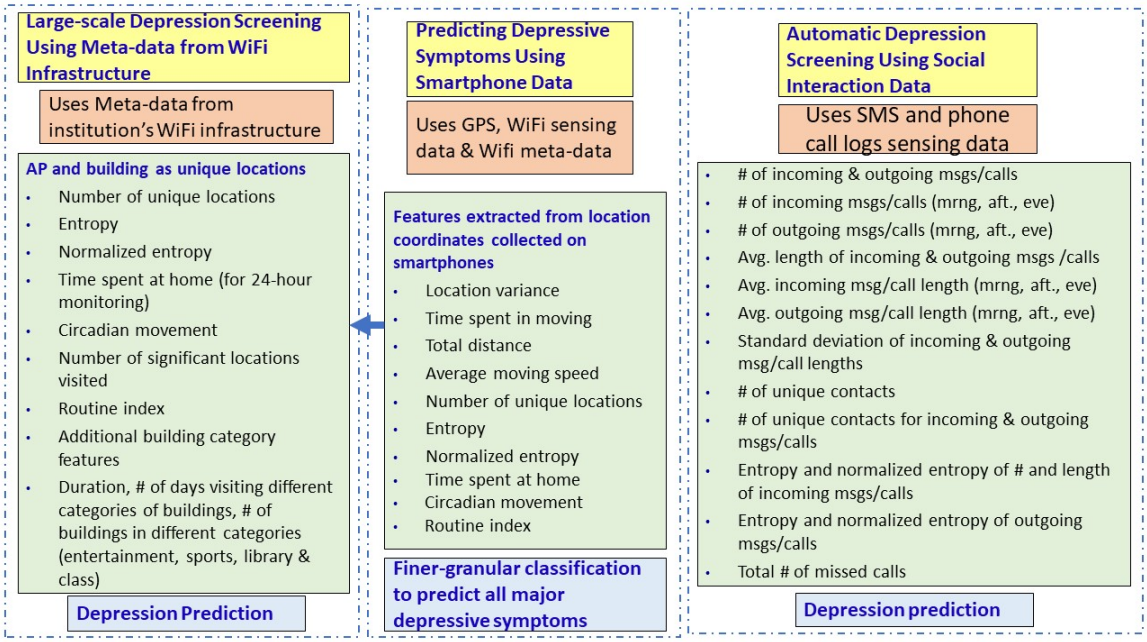


FIGURE 1.2: Overview of the dissertation's problem statements.

Figure 1.2 illustrates an overview of the three problem statements discussed in this dissertation. As shown in the figure, the first problem is to explore a novel approach that uses meta-data from institution WiFi for large scale automatic depression screening. We extract an array of features using access point and building information. In

the second work, we have used smartphone sensing data, particularly location data and WiFi meta-data collected from campus WiFi network, for predicting depressive symptoms. We calculated features based on location coordinates recorded via an app running on smartphones. By considering location information represented by building IDs collected from WiFi meta-data, we calculated similar features like we did for depression screening using WiFi meta-data. In the third problem statement, we have used SMS and phone call logs (collected passively via an app for Android platform) for automatic depression screening. We have calculated a comprehensive set of features using both SMS and call logs. We mainly used the statistical information like frequency, times, number and quantity of messages and calls for feature calculation. Out of the three works, two of them focus on depression prediction, while the other focuses on a finer granularity prediction of depressive symptoms. Also, the first and second works both use location information for prediction depression status and symptoms. Whereas, in the third work, we have used social behavior information for depression prediction.

We next discuss each of the three problem statements. In Section 1.2.1, we present a novel approach that explores the usage of meta-data from WiFi infrastructure for large scale automatic depression screening. Next, section 1.2.2 describes a novel approach to predict depressive symptoms using smartphone sensing data and WiFi meta-data. Section 1.2.3 describes the third problem statement of automatic depression screening using social interaction data particularly SMS and phone call logs.

1.2.1 Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure

Depression is an acute chronic illness that has detrimental health and economic impacts. The current diagnosis methods are burdensome and rely either on physician-administered clinical visits or patient self-administered interview tools. The existing approaches are not only high cost incurring but at the same time also suffer from recall bias.

The existing smartphone app based approaches require running an app in the background continuously to collect sensing data on smartphones which may be burdensome. Additionally, such health monitoring applications that require running an app on phone may not be applicable to large-scale population. It will be limited for individual use as it will require app installation and maintenance. We explore a novel approach that uses meta-data from WiFi infrastructure for large-scale automatic depression screening. The main aim is to develop a depression screening tool that is light weight, low-cost and applicable for large-scale depression screening. Several reasons make this approach an ideal solution for public health intervention as it can be applicable to large-scale population and is available at a lost cost. It also does not require any manual intervention and lastly, as users prefer to connect to institution WiFi network. Furthermore, depression is a chronic disease and analysis is based on data collected over a period of time, so, some missing data may not be critical.

Despite the advantages of this approach, there are some associated challenges. WiFi data is opportunistic, data can only be captured in places with WiFi coverage. Also, WiFi data is of lower resolution than GPS locations collected on phones.

1.2.2 Predicting Depressive Symptoms

With the advancements and achievements in the growing research area of ubiquitous computing, studies have investigated the usage of smartphone sensing data for automatic depression prediction [81, 13, 61, 25]. These studies used smartphone sensing data like GPS, WiFi, activity or WiFi infrastructure meta-data for depression prediction. The main idea is that they extract meaningful features from one or combinations of smartphone sensing data and build a family of machine learning models for depression prediction. However, all the above studies focus on binary classification i.e. predicting whether one is depressed or not.

Depressive symptoms manifest in many aspects of our daily life. A depressive symptom is a finer level aspect of depression status. It can be both behavioral (appetite, energy, psychomotor, sleep disturbance) and cognitive (concentration, interest, self-criticism, feeling sad/depressed). Currently, survey instruments, such as PHQ-9 [44] and Quick Inventory of Depressive Symptomatology (QIDS) [60], are commonly used to detect depression and keep track of the development of the symptoms which are burdensome and difficult to execute on a continuous basis.

We explore a novel approach of investigating the feasibility of using smartphone data (smartphone sensing and WiFi meta-data) for predicting depressive symptoms. Predicting depressive symptoms is at a much finer granular level and provides a detailed picture of an individual's depression conditions. This can be immensely useful for both patients and clinicians.

However, there are challenges associated with prediction of depressive symptoms. Firstly, prediction is at a much finer granular level. Secondly, smartphone data is behavioral in nature e.g. location, activity, while the depressive symptoms are both

behavioral and cognitive.

1.2.3 Automatic Depression Screening Using Social Interaction Data

Most of the existing studies have mainly used location data for depression screening. Some of them merely rely on participants self-reports data as ground truth. There are other critical aspects of human behavior that should be taken into consideration when analyzing a mental health problem like depression. One of such aspects is a person's social life and interaction with others. Social interaction plays an essential role in the day-to-day life of individuals and existing researches has proved that social interaction play a beneficial role in the maintenance of psychological well-being [41]. Researchers have found that both quality and quantity of social relationships affect mental health and mortality risk [53]. SMS and voice calling are most dominant modes of communication among the student population and play a central role in maintaining their social networks [46, 35]. We aimed to explore the feasibility of using the social interaction data using SMS and phone call records collected via an android app from college-age students for depression prediction.

Despite the crucial role of social interaction data on mental health, there are some associated challenges. SMS and phone call data are opportunistic i.e. data will only be recorded if there is a message or a call event. Preference of messaging app is also a limitation. Our analysis uses Android's SMS app, some users might prefer to use other apps. Lastly, this analysis use SMS and phone call data from Android phone users only due to the operating system restrictions of iOS, so we have a comparatively smaller set of users.

1.3 Contributions of This Dissertation

Despite the advantages of the existing depression screening approaches, there are several challenges and drawbacks associated with them. There are numerous directions to explore that probe the possibility of using smartphone data in a much effective way that address the following questions: 1) Is it possible to develop an approach that is light weight, low cost and applicable to a large-scale population for automatic depression screening? 2) Can we predict depression at a much finer granular level of individual depressive symptoms instead of just the depression status i.e. whether one is depressed or not? 3) Can we use social behavior and interaction to predict depression? This dissertation makes contributions that address these questions. Specifically, we present *LifeRhythm* study that developed an automated tool that monitored human behavior through their smartphones for automatic depression screening. We used both smartphone sensing data and meta-data from WiFi infrastructure for this purpose. We developed a *LifeRhythm* app [25]. The app runs in the background, passively collecting sensing data with no need of user interaction. The Android version of the app was developed based on an existing publicly available library, Emotion Sense library [45]. This app records a wide array of smartphone sensing data like GPS, Wifi, activity, SMS, phone call records and others. For iPhone, the app was developed using Swift from scratch.

Also, the results are based on a comparatively larger data set that is collected from a two phase study. Each study lasted for several months including tens of participants recruited from the University of Connecticut. Larger scale of study allowed us to construct a family of machine learning models keeping into consideration the variations among human beings. Considering larger population helps in providing

results that solve depression as a public health problem. Furthermore, all the three works discussed in this dissertation use the data that was passively collected without any efforts from the users.

Automatic depression prediction using the collected data provides a strong foundation of keeping track of depression and depressive symptoms. Sections 1.3.1, 1.3.2 and 1.3.3 present the contributions of this dissertation for automatic depression prediction and prediction of depressive symptoms respectively.

1.3.1 Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure

The following are the key contributions of our research when using meta-data from WiFi infrastructure for large scale automatic depression screening:

- We present a novel approach that uses the meta-data collected from WiFi infrastructure for large-scale automatic depression screening. The main intuition is that when a phone connects to a near-by access point (AP), the location of the AP can be used to infer an approximate location of the users. We construct a family of machine learning models using the Wifi association records collected by the IT services at the University of Connecticut for predicting depression status. This approach does not require running an app on smartphones.

- The results show that WiFi association data collected passively from the campus WiFi network is effective for depression screening. For predicting depression (i.e., classifying whether one is depressed or not), the F_1 scores are as high as 0.85, comparable to those obtained using data collected by instrumenting smartphones [61, 25, 87].

- We find that building based features have stronger correlation with self-report scores than AP based features, and lead to better classification results than using AP based features. Furthermore, including building category features further improves the classification results.

- Using behavioral data from the WiFi association records, we have constructed multi-feature regression models to predict PHQ-9 and QIDS scores. We observe that the multi-feature models, in particular, ℓ_2 regularized non-linear models, can significantly improve upon the models that use a single feature for prediction. The correlation between the regressed values the ground-truth values is in a similar range as that obtained when using data directly from phones [61, 25, 87].

- We envision two deployment scenarios for the models constructed from this approach. One for depression screening at the population level, and the other for individual users. At the population level, for instance, after certain new policies or facilities have been established in an institution (e.g., building a gym, establishing a mental clinic), the population level statistics can be useful to assess the effectiveness of these new policies or facilities. As another example, for a university with multiple regional campuses, the population level statistics can be helpful to understand which campus is better in terms of students' mental health and why. At an individual level, a user may elect to use the service to automatically monitor his/her conditions, and receive the results periodically or on-demand. Such results will be immensely helpful for clinicians to make effective treatment decisions.

1.3.2 Predicting Depressive Symptoms

We used smartphone sensing data particularly location data and WiFi infrastructure meta-data for predicting depressive symptoms. The following are the contributions of this dissertation for predicting depressive systems:

- We predict all major categories of depressive symptoms automatically using smartphone data. For this purpose, we used smartphone sensing data collected by running an app on smartphone and meta-data collected from the WiFi infrastructure.

- The prediction is done at a much finer granular level instead of binary classification i.e. predict if a person is depressed or not.

- The prediction of depressive symptoms provides a detailed picture on depression conditions that could be tremendously helpful to both clinicians and patients.

- We find that sensing data collected directly on smartphones can predict a rich set of depressive symptoms accurately, including both behavioral (appetite, energy, sleep, psychomotor) and cognitive symptoms (interests, self-criticism, feeling depressed, concentration). The predicted F_1 scores can be as high as 0.83, comparable to the F1 scores obtained for predicting the overall depression status [25, 87, 49, 89]. In addition, we observe stronger prediction results for the depressed participants compared to the non-depressed participants.

- We find that meta-data collected from institution's WiFi infrastructure can also predict a variety of depressive symptoms accurately. Specifically, we explore 24-hour monitoring (for the users who spend time during both night and day

on campus, e.g., those who live on campus), and daytime monitoring where only the daytime information (8am-6pm) is available (e.g., for those who are only on campus during daytime). We find that even daytime information is sufficient to provide accurate prediction for a set of depressive symptoms. Our results demonstrate that the meta-data collected from an institution's WiFi infrastructure can be used to keep track of the wellness of a large population at very little cost.

- We further explore predicting finer-level depressive symptoms, e.g. increased or decreased appetite/weight, feeling restless or slowed down, and sleep disturbance (time taken falling asleep, sleep during night, sleeping too much, and waking up too early). Our results demonstrate that even finer-level depressive symptoms (particularly sleep related) can be predicted accurately using smartphone data, with predicted F_1 scores up to 0.86.

- Our results show that both behavioral and cognitive symptoms can be predicted accurately using smartphone data (that is behavioral in nature).

1.3.3 Automatic Depression Screening Using Social Interaction Data

The following are the key contributions of our research work when we used social interaction data particularly SMS and phone call records for depression prediction:

- Our study makes a significant contribution by using an array of features based on different categories of SMS and call usage patterns and diversity. They are calculated based on statistical information carried by SMS and phone call data

for depression prediction. Our results show that each of the sources (SMS and phone call) can independently predict depression effectively.

- Our results show that SMS data can predict depression accurately. The highest predicted F_1 score is obtained from XGBoost model and is as high as 0.80. We also found that majority of the depressed individuals have higher number of incoming and outgoing messages but to a specific group of contacts.

- We find that phone call records can effectively predict depression. The best model is XGBoost and predicted F_1 score is as high as 0.78. Additionally, we also observed that individuals with depression spend more time on phone calls as compared to the non-depressed population.

1.4 Dissertation Roadmap

The dissertation is structured as follows. The first part of the dissertation (Chapter 2) presents our novel approach on large-scale automatic depression screening using meta-data from WiFi infrastructure. High-level approach of the system and data analysis methodology are discussed in detail. Next, results of the analysis of WiFi meta-data performed at three levels: Access point (AP), building level and enhanced building level analysis are discussed in detail in this chapter. In the second part of the dissertation (Chapter 3), prediction of depressive symptoms using smartphone data is presented in detail. This chapter describes high-level approach followed by the results from analysis using smartphone sensing data and WiFi infrastructure meta-data. Furthermore, analysis related to prediction of finer-level depressive symptoms is presented. In the third part of the dissertation (Chapter 4, our work on automatic

depression prediction using social interaction data is presented. High-level approach followed by data preprocessing methodology is discussed. Depression prediction results for various time windows used in the analysis are discussed in detail. In the final part of the dissertation (Chapter 5), I conclude my dissertation with the insights received from the observations and results of the research works. Next, future work is presented that discusses other future directions of the applications of smartphone sensing for mental health and wellness that need to be unleashed and explored.

Chapter 2

Large-scale Automatic Depression Screening using Meta-data from WiFi Infrastructure

2.1 Introduction

Depression is a common mental health problem that affects 350 million people worldwide [78]. It has serious consequences on both physical and psychological functioning. People with depression suffer from higher medical costs, exacerbated medical conditions, and much higher mortality [67, 40, 20]. Suicide rate due to depression has tremendously increased in the past several years [78]. Reports published in 2010 show that in the United States, suicide is the 10th leading cause of death, and 70% of these suicide victims are reported to have a mood disorder such as depression [1]. Diagnosis of depression has been based on physician-administered or patient self-administered interview tools [70], which are burdensome and difficult to carry out on a continuous

basis. In addition, responses to these tools are often subjective (depending on a user's current mood) and limited by recall bias.

The ubiquitous adoption of smartphones has presented new opportunities for depression screening. Several recent studies (e.g., [81, 61, 13, 25], see details in Section 4.2) have proposed novel approaches that use smartphones for automatic depression screening. The intuition of these approaches is that, since smartphones are equipped with a rich set of sensors (e.g., GPS, activity, light) and are constantly carried by their owners, they can be used as effective "human sensors" for cataloging many aspects of their users' behavior. Such behaviorial features can then be fed into machine learning algorithms (with pre-trained machine learning models) to automatically detect depression. All existing approaches, however, require running a mobile app on users' phones, which continuously captures various sensing information on the phones.

In this part of the dissertation, we explore a novel alternative approach that requires no direct data capture on a user's phone. Instead, it uses WiFi association meta-data that are collected passively from an institution's WiFi network (e.g., the campus WiFi network of a university, company or military base). The rationale is as follows. WiFi networks have been deployed widely by institutions as a convenient wireless communication infrastructure. Once connected to the WiFi infrastructure, the locations of a smartphone (and hence the user) can be roughly determined by the access points (APs) that it is associated with (a phone must associate with a close-by AP for Internet access). Therefore, the AP association records of the WiFi infrastructure can be used to infer the locations of the users over time; these location transcripts can be used for depression screening.

The above approach does not require installing app or collecting data directly

from individual smartphones. Instead, it leverages WiFi association data that can be easily collected (and indeed are routinely collected in many institutions for network management and diagnosis), and can provide large-scale depression screening for thousands of users simultaneously at very little cost, making it an ideal approach for public health intervention (see discussion on usage of the data and user privacy considerations in Section 2.3.2). On the other hand, compared with the approaches that use sensing data collected on the phones, this approach has to contend with two challenges: (i) the location data is of lower resolution: an AP association event only indicates that a user is close to the AP, which is of lower resolution compared to GPS locations collected on phones. (ii) the data collection is opportunistic, since the locations can only be captured when a phone is connected to the WiFi infrastructure.

To explore the feasibility of the above approach, we have analyzed two datasets, collected during Phase I and Phase II of our study, respectively. Each study lasted for several months, including tens of participants recruited from a research university in the US. We consider two scenarios: one for the participants who spend time during both night and day on campus and hence yield meaningful data over the full 24 hour period each day; the other only considers daytime (8am-6pm) data, corresponding to the commuting scenario, where participants are only present on campus during daytime. For both studies, we have analyzed the WiFi association meta-data collected from the campus WiFi network, direct assessment by a clinician, and the participants' self-reports, specifically, Patient Health Questionnaire (PHQ-9) [44] for Phase I and Quick Inventory of Depressive Symptomatology (QIDS) [60] for Phase II, that were collected periodically over time.

Our analysis is at three levels: AP level, building level, and enhanced building level. In AP level analysis, we treat each AP as a unique location, while at the

building level, we treat all the APs that are in the same building as the same location. The enhanced building level analysis further enhances the building level analysis by including additional building category related features to infer the activity of a user. We make the following main contributions:

- Our results show that WiFi association data collected passively from the campus WiFi network is effective for depression screening. For predicting depression (i.e., classifying whether one is depressed or not), the F_1 scores are as high as 0.85, comparable to those obtained using data collected by instrumenting smartphones [61, 25, 87].

- We find that building based features have stronger correlation with self-report scores than AP based features, and lead to better classification results than using AP based features. Including building category features further improves the classification results.

- Using behavioral data from the WiFi association records, we have constructed multi-feature regression models to predict PHQ-9 and QIDS scores. We observe that the multi-feature models, in particular, ℓ_2 regularized non-linear models, can significantly improve upon the models that use a single feature for prediction. The correlation between the regressed values the ground-truth values is in a similar range as that obtained when using data directly from phones [61, 25, 87].

The rest of the chapter is organized as follows. Section 4.2 describes related work. Section 3.2 outlines our high-level approach and discusses deployment issues. Section 2.4 describes data collection. Sections 2.5, 2.6 and 2.7 present our analysis

at the AP level, building level, and enhanced building level, respectively. Finally, Section 4.7 concludes the chapter and briefly describes future work.

2.2 Related Work

Recent studies have used sensing data collected from smartphones for detecting depression or depressive mood [28, 31, 32, 13, 81, 61, 6, 50, 91, 80, 54, 24, 25, 71, 19, 87, 49]. Wang et al. [81] studied the impact of workload on stress and day-to-day activities of students. They found significant correlation between the behavioral traits (in terms of conversation duration, number of locations visited, sleep) and depressive mood. Saeb et al. [61] found significant correlation between the phone usage and mobility patterns with respect to the self-reported PHQ-9 scores. Canzian and Musolesi [13] studied the relationship between the mobility patterns and depression, and found that individualized machine learning models outperformed general models. Farhan et al. [25] found that the features extracted from the smartphone sensing data can predict depression with good accuracy. Yue et al. [87] investigated fusing GPS and WiFi association data, both collected locally on smartphones, for more complete location information for improved depression detection. Lu et al. [49] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. Suhara et al. [71] developed a deep learning based approach that forecasts severely depressive mood based on self-reported histories. All the above studies use sensing data collected directly from smartphones, which requires installing an app on the phones. Our study investigates an alternative approach that uses large-scale data collected directly from a WiFi infrastructure. These two ap-

proaches present different strengths and weaknesses (see Section 2.3.3). One main contribution of this work is that we investigate the feasibility of the WiFi infrastructure based approach, and demonstrate that it can achieve comparable performance for depression screening as the approach based on instrumenting smartphones.

There is a rich literature on analyzing WiFi data. The focus has been primarily on the aspects of networking and data communication, with a few studies on inferring user behaviors. For instance, the studies in [37, 38] used WiFi traffic to mine user behavior patterns (e.g., identify behavior groups). The study in [36] proposed a system that discovers social interaction based on opportunistic probe request and null data frames sent by mobile devices. The wellness monitoring platform proposed by [79] used employee's everyday devices and existing infrastructure (interconnected desktop/laptop, enterprise WiFi) for activity tracking and physiological measurements (i.e., heart rate). Their system was proposed to reduce potential health risks associated with prolonged sitting in office environments. In addition, existing research has leveraged WiFi access data for studying geospatial activity and user behavior [90], mental state, including depression [6], and population-level monitoring [39]; these studies, however, used the WiFi data collected by the phones, not by the WiFi infrastructure.

To the best of our knowledge, our study is the first that uses WiFi meta-data collected from institution WiFi infrastructure for depression screening. Our approach does not require instrumenting smartphones to collect data. It can be particularly beneficial for public health intervention in an institution (e.g., university, military base, or company).

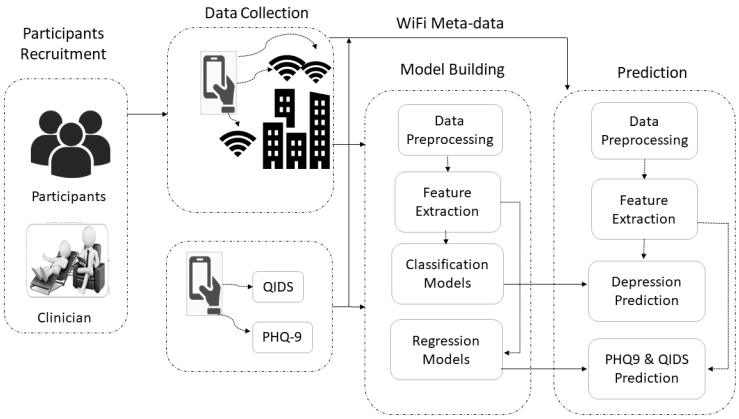


FIGURE 2.1: Illustration of our high-level approach.

2.3 High-level Approach and Deployment Considerations

2.3.1 Background

By virtue of their untethered nature, ease of setup, and mobility support, WiFi networks have been widely deployed in institutions (e.g., universities, companies) as a wireless communication infrastructure. To provide dense coverage, typically multiple access points (APs) are installed on each floor of a building; a user associates with a close-by AP for Internet access. While both cellular and WiFi networks are commonly used for wireless Internet access, whenever available, people often prefer to connect to the WiFi infrastructure since it is free, requires less energy, and has high bandwidth [5, 22, 43, 56, 4]. When connected to a WiFi network, the location of a smartphone can be roughly determined by the AP that it is associated with. Therefore, AP association records collected from WiFi infrastructure can provide location information of a user over time, which can be useful to infer user behaviors

for depression screening.

2.3.2 Model Building, Deployment, and Privacy Considerations

Figure 2.1 illustrates our high-level approach. As shown in the figure, our approach contains two stages: learning prediction models and using the models for prediction. In the first stage, we recruit a population of study participants, and collect their anonymized WiFi network meta-data (i.e., WiFi association records, which indicate the locations of users) along with regular self-reports (e.g., responses to PHQ-9 or QIDS questionnaires) and the results of clinical interviews at a secure server. High-level features are extracted from the data, which are then used to train a family of models to predict the self-report scores and depression status.

In the second stage, the models learned from the first stage will be used for predicting depression. We envision two deployment scenarios: one for depression screening at the population level, and the other for individual users. At the population level, anonymized WiFi network meta-data of the users need to be collected, and fed into the pre-learned prediction algorithms to detect depression. The prediction models can be used to estimate the rate of depression at the population level, which can be used for multiple applications. For instance, after certain new policies or facilities have been established in an institution (e.g., building a gym, establishing a mental clinic), the population level statistics can be useful to assess the effectiveness of these new policies or facilities. As another example, for a university with multiple regional campuses, the population level statistics can be helpful to understand which campus is better in terms of students' mental health and why. At an individual level,

a user may elect to use the service to automatically monitor his/her conditions, and receive the results periodically or on-demand. In this case, while certain identity information needs to be kept so that a user can retrieve his/her information later on, user privacy can be preserved through cryptographic techniques. One approach is private information retrieval [17, 18], where a server keeps track of the prediction results for a set of users, and the information is retrieved so that the server is not aware of what information a user retrieves. Several state-of-the-art protocols (e.g., [29, 3]) can be used for this purpose.

As with any work that applies machine learning to passively collected data, user privacy and responsible usage of the data are important considerations in the system design, implementation and deployment. For both population and individual level deployment (as outlined above), an institution needs to carefully design and implement the mechanisms for user consent and preserving user privacy. For population level deployment, no user identity needs to be kept; for individual level deployment, the collected data and predicted results need to be associated with certain form of identity information for later retrieval, and hence carries even more privacy implications. Detailed design, implementation and deployment mechanisms are beyond the scope of this work. Instead, our focus is on exploring the feasibility of learning accurate prediction models using WiFi meta-data.

2.3.3 Pros and Cons of the Proposed Approach

Compared to the approaches that use sensing data collected on the phones (by running an app on a phone), our approach of using WiFi meta-data from an institution's WiFi infrastructure has both advantages and disadvantages. The most salient advantage

is probably the large-scale data that can be used for depression prediction at the population level (for an institution), which is difficult to achieve when instrumenting individual phones. Another (arguably) advantage of the WiFi infrastructure based approach is that the data can be easily collected through standard network protocols (indeed many institutions routinely collect such data for network management and diagnosis), without the need of designing, installing and running an app on individual phones. On the other hand, collecting data using an institution's WiFi infrastructure for depression screening needs buy-ins from the institution. In addition, as mentioned earlier, it needs carefully designed and executed mechanisms for data protection and consenting process, which can be more difficult than the corresponding tasks when collecting data on individual phones (which can simply store the data and run the prediction models locally on the phone).

Two disadvantages/challenges of the WiFi infrastructure based approach are: (i) the location data is of lower resolution than GPS locations collected on phones, and (ii) the collected data is opportunistic—the locations can only be captured in places with WiFi coverage (e.g., indoors) and when the WiFi connection of a phone is active (a phone may duty cycle its WiFi connection to preserve energy). We anticipate that, despite the above two limitations, the data from WiFi infrastructure can still provide a valuable overview of a user's behavior. This is because, as mentioned earlier, whenever available, users prefer to connect their smartphones to WiFi networks due to performance and cost considerations. Furthermore, after choosing a WiFi network for Internet access, most smartphones will periodically connect to the network, so that different background services (e.g., email or Facebook client) can get updates. In addition, given that depression is a chronic disease, the detection can be based on data collected over a period of time, and occasional missing data may not be a

critical limitation.

2.3.4 Data Analysis Methodology

The rest of the chapter focuses on exploring whether the data from WiFi infrastructure can be used for effective depression screening, despite its coarse-grain and opportunistic location data collection. We will investigate three approaches for analyzing WiFi meta-data: the first is the AP level analysis, the second is the building level analysis, and the third enhances the second by adding more building semantics information. Specifically, the first approach simply treats each AP as a unique location, and investigates the characteristics of the locations that a user visits during a time period. It is simple, requiring no detailed information of the APs. The second approach treats each building as a unique location. As such, it requires knowing which building an AP is located in, and treating an association event to an AP as a visit to the corresponding building. The third approach further uses the category of a building (the category is based on the main purpose of the building, e.g., entertainment, sports, library, or classroom building) to infer potential activity of a user. It therefore requires even more information (knowing the main purpose of the buildings and classifying the buildings into the corresponding categories).

The first approach (AP level analysis) uses the least amount of information, and serves as a baseline. The second approach (building level analysis) uses more information (i.e., mapping APs to the buildings). It uses coarser-grain location information (since it does not differentiate the APs in the same building), but intuitively may represent the locations in a more semantically meaningful way. The third approach (enhanced building level analysis) uses the most information out of the three ap-

proaches, and serves to investigate whether adding more semantic information of the buildings leads to better performance.

For each of the above three approaches, we further consider two scenarios: one using data collected over 24 hours each day, covering both daytime and nighttime location information, and the second only uses data collected during daytime (8am-6pm). The first is applicable to the scenario where a user spends significant amount of time during both night and day on campus (e.g., a student living in a dorm on campus), while the second corresponds to a commuting scenario, where an employee (or student) comes to a company (university) for work (study) during daytime, and then spends the rest of the time off campus. Clearly, 24-hour data provides much more insights into a user's behavior. We are also interested in the second scenario to investigate whether daytime location information alone is sufficient to detect depression. Existing approaches that collect data directly from smartphones belongs to 24-hour monitoring, since they collect data continuously during both daytime and nighttime.

2.4 Data Collection

Our study was conducted in the University of Connecticut. The study was in two phases: Phase I and Phase II, both approved by the university's Institutional Review Board (IRB). Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017. For both phases, the participants were full-time students of the university, aged 18-25. We recruited 79 participants in Phase I study. Of them, 73.9% were female and 26.1% were male. In terms of

ethnicity, 62.3% were white, 24.6% were Asian, 5.8% were African American, 5.8% had more than one race and 1.5% were other or unknown. For Phase II study, we recruited 103 participants (76.7% female and 23.3% male; 58.25% white, 25.24% Asian, 3.88% African American, 7.77% having more than one race and 4.85% being other or unknown). All participants met with our study clinician for informed consent and initial screening before being enrolled in the study.

Based on the clinician assessment, in Phase I study, 19 participants were classified as depressed and the remaining 60 participants were classified as non-depressed; in Phase II study, 39 participants were classified as depressed and the remaining 64 participants were classified as non-depressed. In both cases, our recruitment intended to recruit the same number of depressed and non-depressed participants, and was not able to recruit as many depressed participants as intended.

Each participant used a smartphone to participate in the study. Their phones were configured so that they connected to the university's campus WiFi network as the default method to access the Internet. We recorded the MAC addresses of their phones, which were hashed to 16 bytes for anonymity, and used later on to identify their corresponding records in the WiFi association data (see Section 2.4.1). In addition, each participant used an app that we developed to fill in PHQ-9 questionnaire (for Phase I) or QIDS questionnaire (for Phase II) periodically, which was encrypted and sent to a secure server. To ensure the privacy of participants, we assigned a random ID to each participant, which was used to identify the participants. Three types of data were collected: WiFi association data, questionnaire responses from the participants, and clinician assessment. We next describe these data in more details.

2.4.1 WiFi Association Data

The WiFi association data were collected by the university's IT services. They were sent to us on a regular basis. Each record corresponds to an AP association event, represented as a tuple (a_i, u_i, t_i, d_i), where i is the row index for the event in the dataset, a_i is the MAC address of an AP, u_i is the MAC address of a wireless device, t_i is the start time, and d_i is the duration of the association event. This tuple indicates that the device (and hence the user) was close to the location of a_i during $[t_i, t_i + d_i]$. For building level analysis, we further use additional information provided by the university IT services to determine the building that each AP is located in, and regard that the device (and the user) is in the corresponding building during $[t_i, t_i + d_i]$. We further classify the buildings on campus into multiple categories, including entertainment (e.g., in theatre, performing arts center), sports (e.g., in student recreation facilities), library, class (i.e., classroom buildings), and others. These categories are then used to extract features related to a particular category of buildings (see Section 2.7.1). To preserve user privacy, for each AP association record, we hashed the MAC address of the AP to anonymize it (in the same way as we hashed the participants' MAC addresses), and only stored the anonymized data on the server. The AP association data of the participants were retrieved based on their hashed MAC addresses. Since most students were not on campus during the holidays (Thanksgiving and Christmas) and breaks (spring, winter and summer breaks), our data analysis below excluded those time periods.

2.4.2 Questionnaire Responses

In Phase I study, participants were asked to fill in PHQ-9 Questionnaire [44] every two weeks. PHQ-9 is a 9-item questionnaire that is self-reported by the users. Clinicians use it to diagnose and monitor depression. Every question in PHQ-9 asks a person's mental and behavioral state in the past two weeks (which is why we asked a participant to fill in the questionnaire every two weeks). PHQ-9 scores are calculated based on a person's answer for each question. The minimum score is 0 and the maximum score is 27. A participant filled in a PHQ-9 questionnaire during the initial assessment, and then on her (his) phone every 14 days. Reminders to users were sent three days after their PHQ-9 filling due date if we missed their reports.

In Phase II study, following the suggestion of our study clinician, we switched from PHQ-9 to a more comprehensive questionnaire, QIDS [60]. The reason for switching to QIDS is two-fold. Firstly, QIDS provides more detailed information than PHQ-9, and hence allows finer-grained labeling of depression symptoms. For instance, instead of asking a single question on decreased or increased appetite as in PHQ-9, it differentiates these two types of appetite changes. As another example, QIDS asks four questions regarding sleep, instead of a single question in PHQ-9. Secondly, the frequency of QIDS is every week (each question in QIDS asks a participant to reflect on the past week), which allows us to obtain more frequent self-reports from participants. As PHQ-9, QIDS is also widely used in clinical settings. It measures 16 factors across 9 different criterion domains including 1) mood, 2) concentration, 3) self-criticism, 4) suicidal ideation, 5) interests, 6) energy/fatigue, 7) sleep disturbance (initial, middle, and late insomnia or hypersomnia), 8) decrease or increase in appetite or weight, and 9) psychomotor agitation or retardation. The total score ranges from

0 to 27. A participant filled in a QIDS questionnaire during the initial assessment, and then on her (his) phone every 7 days.

As we shall see (Sections 2.5 to 2.7), the different self-report instruments used in Phases I and II studies lead to different correlation and regression results. On the other hand, the classification results for Phases I and II are similar.

2.4.3 Clinical Assessment

Each participant in the study had an initial screening with our study clinician. The clinician classified a participant as depressed or non-depressed following a Diagnostic Statistical Manual (DSM-V) based interview and the participant's PHQ-9 or QIDS evaluation. A depressed participant must be in treatment to remain in the study. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician). Each meeting lasted 10-20 minutes and only involved interviews to assess psychiatric symptoms. The purpose of the interviews was to correlate and confirm their self-reported PHQ-9 or QIDS scores with their verbal report.

2.5 AP level Analysis

In this section, we present our results on AP level analysis. Specifically, we treat each AP as a unique location; if a WiFi association record indicates that a user is associated with an AP a from time t to t', then we regard that the user is at location a from t to t'. In the following, we first present our data preprocessing procedure, and then describe feature extraction and analysis results.

2.5.1 Data Preprocessing

As mentioned earlier, for both Phase I and Phase II studies, we consider two scenarios: 24-hour monitoring and daytime monitoring. The first scenario only considers the users who spent time during both night and day on campus. Since all the participants were university students, they naturally spent time on campus during the day, but they might not spend time on campus during nighttime (e.g., the commuting students). We therefore identify users for the first scenario as those who spent at least 40% of the time (chosen empirically) on campus during 12-6am (typically corresponding to sleeping time), as indicated by the WiFi association records. These participants likely lived on campus (we did not collect information on whether a user lived on campus or not, and were not able to verify whether this was indeed the case). The second scenario considers all the users.

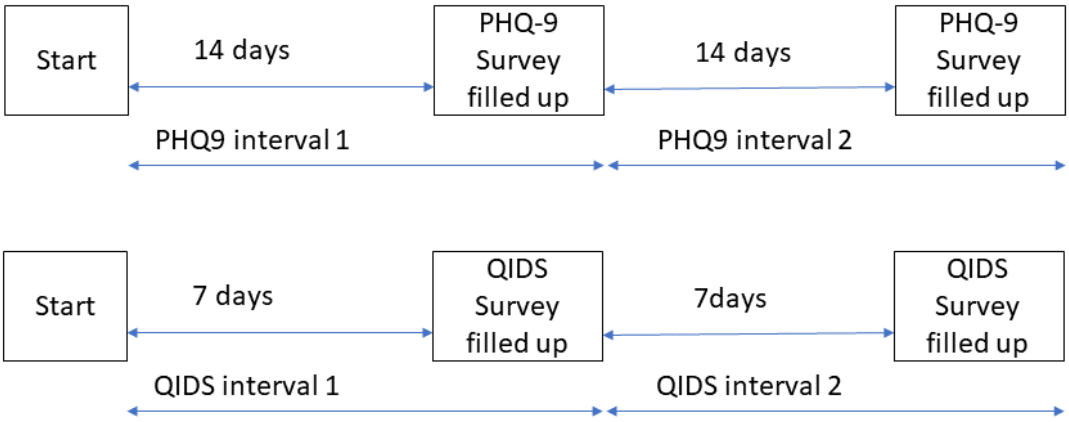
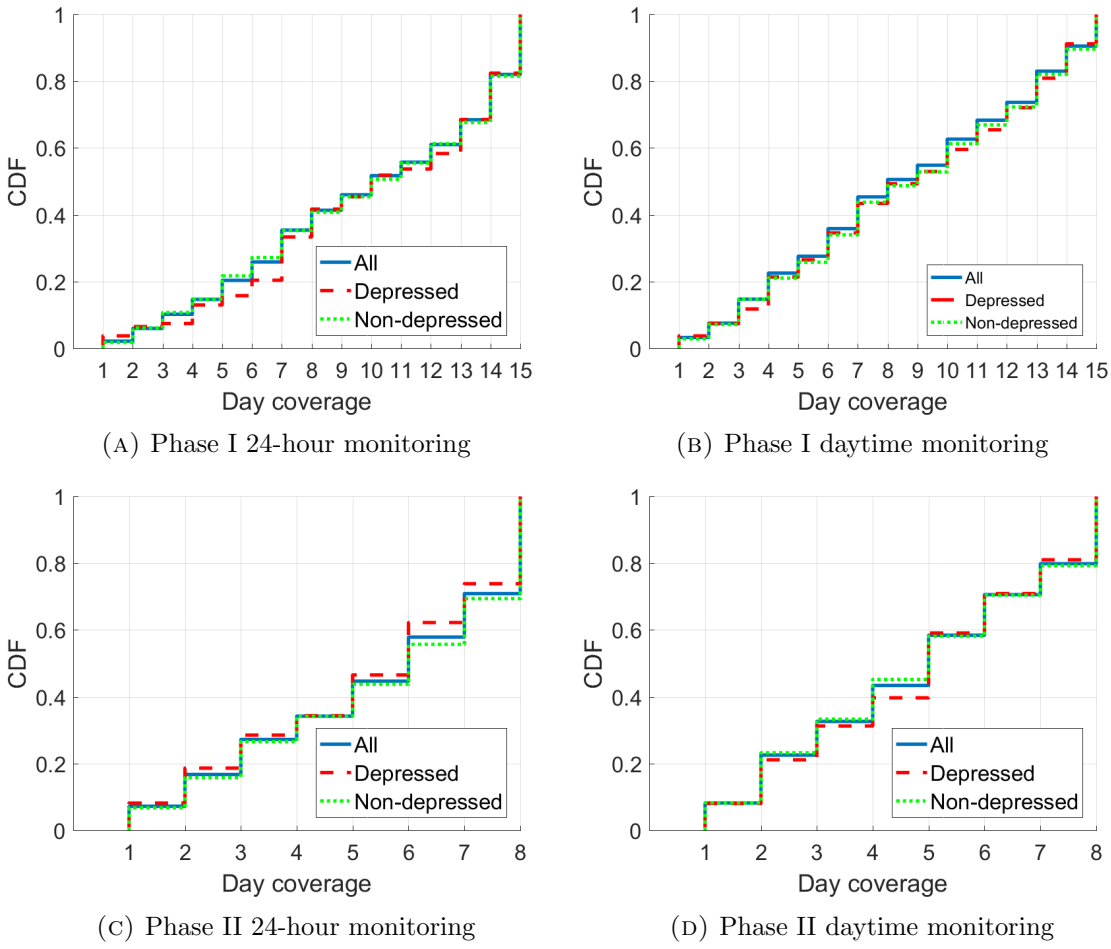


FIGURE 2.2: Illustration of PHQ-9 and QIDS intervals.

Phase I data preprocessing. In Phase I study, a user was asked to fill in a

PHQ-9 questionnaire as a self-report every 14 days. We define a *PHQ-9 interval* as a 15-day time interval, including the day when a user fills in a PHQ-9 questionnaire and the previous 14 days, as illustrated in Figure 2.2. For each participant, we have organized the data collected for each PHQ-9 interval, and mapped it with the corresponding PHQ-9 score.



(A) Phase I 24-hour monitoring

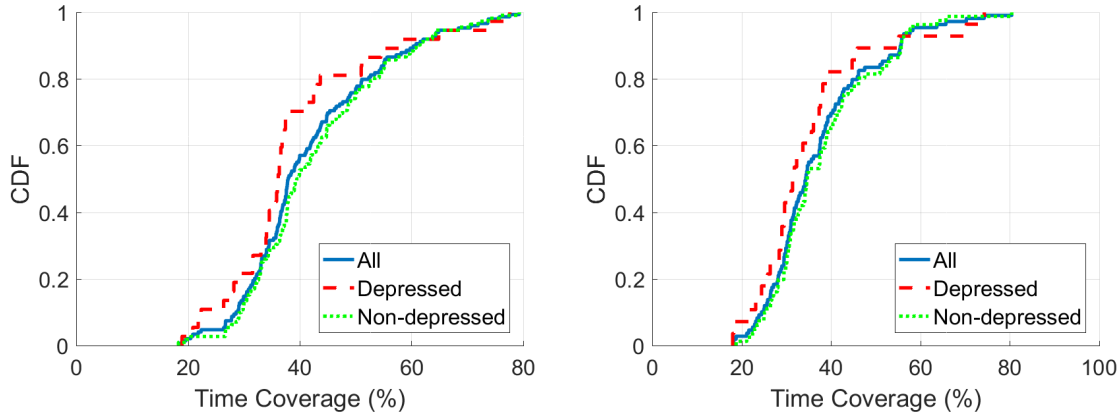
(B) Phase I daytime monitoring

(C) Phase II 24-hour monitoring

(D) Phase II daytime monitoring

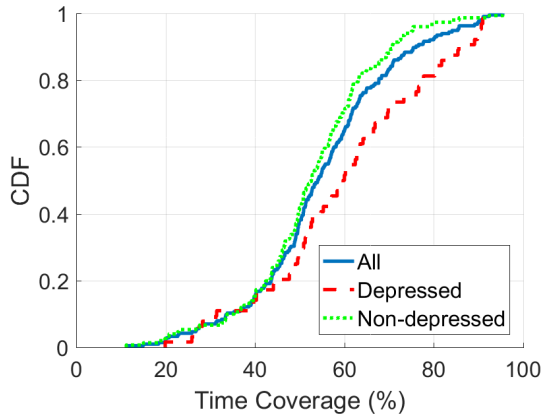
FIGURE 2.3: Day coverage of the campus WiFi meta-data for various scenarios.

Figure 2.3(A) plots the cumulative distribution function (CDF) of the day coverage (i.e., the number of days with WiFi association data) of the PHQ-9 intervals for 24-hour monitoring. The results for three cases, all participants, depressed participants

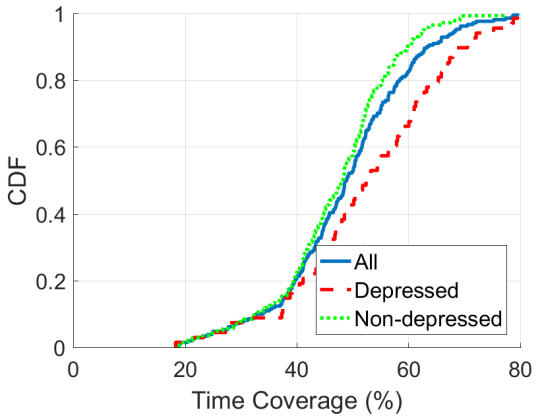


(A) Phase I 24-hour monitoring

(B) Phase I daytime monitoring



(C) Phase II 24-hour monitoring



(D) Phase II daytime monitoring

FIGURE 2.4: Time coverage of the campus WiFi meta-data for various scenarios.

only, and non-depressed participants only, are plotted in the figure. Figure 2.3(B) plots the corresponding results for Phase I daytime monitoring. We see that the day coverage varies from 1 to 15 days. The reason for not capturing any WiFi association data during a day might be due to multiple reasons, e.g., the malfunction of the WiFi data capture equipment, a user not being on campus, a user turning off the WiFi interface on the phone, or a phone being out of battery. To deal with the missing data, we only include the PHQ-9 intervals that contain at least d days of data in our analysis. We set d to 12, 13 or 14. The results below are based on the most conservative threshold, i.e., when $d = 14$ (the prediction results for the other two thresholds are similar, and are omitted in the interest of space). In addition, to exclude the cases when a user just passed by an AP (without staying at the location), for a PHQ-9 interval, we only consider those APs where a participant spent at least 15 minutes over the PHQ-9 interval.

After the above data preprocessing procedures, for Phase I 24-hour monitoring, we obtained a total of 149 intervals, accounting for 31.6% of the total number of intervals for this case (which were from a subset of 47 participants who spent time during both daytime and nighttime on campus). Out of these, 37 belonged to depressed participants and 112 belonged to non-depressed participants. A total of 37 users were found, with 11 depressed and 26 non-depressed. For Phase I daytime monitoring, we obtained a total of 109 PHQ-9 intervals, accounting for 16.4% of the total number of intervals for this case (which were from all participants in Phase I), with 28 belonging to depressed participants and 81 belonging to non-depressed participants; 35 users were found, with 10 identified as depressed and 25 as non-depressed.

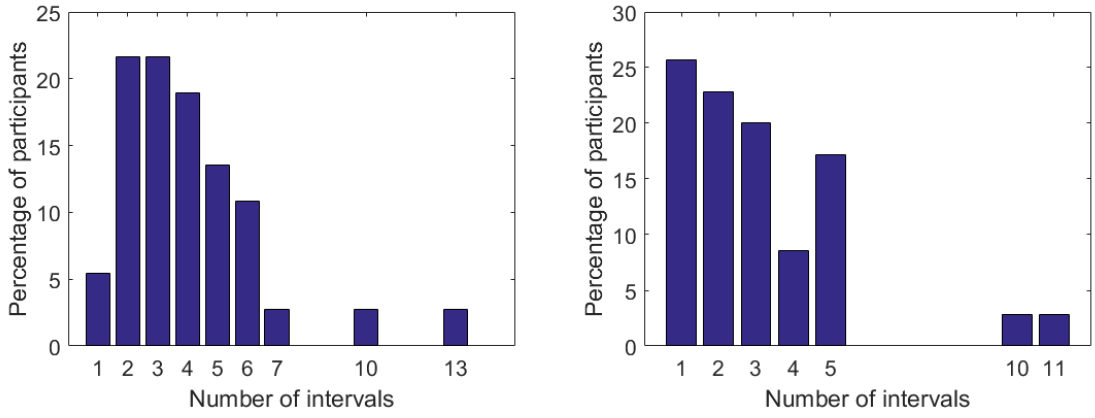
Figure 2.4(A) plots the CDF of the time coverage (i.e., the percentage of time with WiFi association data during a PHQ-9 interval) for 24-hour monitoring. We see

that the time coverage varies from 20% to 80%. As mentioned earlier, since the data capture is opportunistic, the time coverage varies, and only around 30% of the PHQ-9 intervals have time coverage above 50%. We observe similar results for daytime monitoring, as shown in Figure 2.4(B).

Phase II data preprocessing. In Phase II study, a user was asked to fill in a QIDS questionnaire every 7 days. We define a *QIDS interval* as a 8-day time interval including the day when a user fills in a QIDS questionnaire and the previous 7 days (illustrated in Figure 2.2). Figures 2.3(C) and (D) plot the CDF of the day coverage for the QIDS intervals for 24-hour and daytime monitoring, respectively. We see that the day coverage varies from 1 to 8 days; around 20-30% of the QIDS intervals have the maximum day coverage of 8 days. We considered three scenarios, where we only included the QIDS intervals that contain at least 6 or 7 days of data in our analysis. The results below are based on the more conservative threshold, i.e., using the QIDS intervals that contain at least 7 days of data. Again, to exclude the cases when a user just passed by an AP, for a QIDS interval, we only consider those APs where a participant spent at least 10 minutes over the QIDS interval.

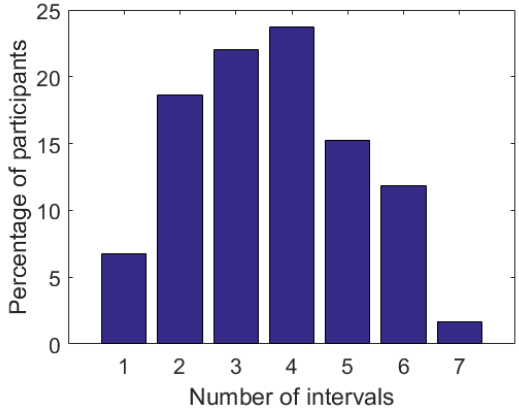
After the above data filtering, for Phase II 24-hour monitoring, we extracted a total of 215 QIDS intervals, accounting for 41.3% of the total number of intervals for this case (which were from a subset of 66 participants who spent time during both daytime and nighttime on campus). Among them, 64 and 151 intervals belong to depressed and non-depressed participants, respectively. These data belonged to a total of 59 users, with 19 as depressed and 40 as non-depressed. For Phase II daytime monitoring, we extracted 211 QIDS intervals, accounting for 28.3% of the total number of intervals (i.e., from all participants in Phase II), with 68 and 143 intervals belonging to depressed and non-depressed participants, respectively; these data belonged

to 74 users, with 26 as depressed and 48 as non-depressed. Figures 2.4(c) and (d) plot the CDF of the time coverage for 24-hour monitoring and daytime monitoring, respectively. The time coverage varies from 10% to 90%.

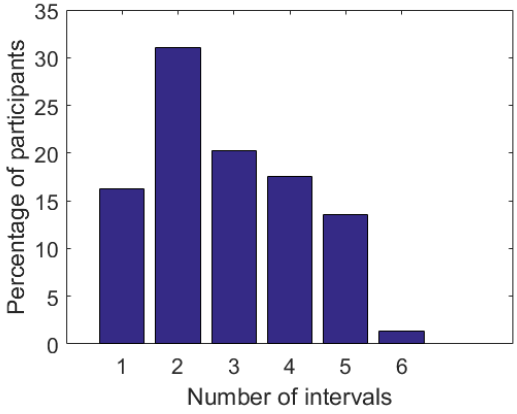


(A) Phase I 24-hour monitoring

(B) Phase I daytime monitoring



(C) Phase II 24-hour monitoring

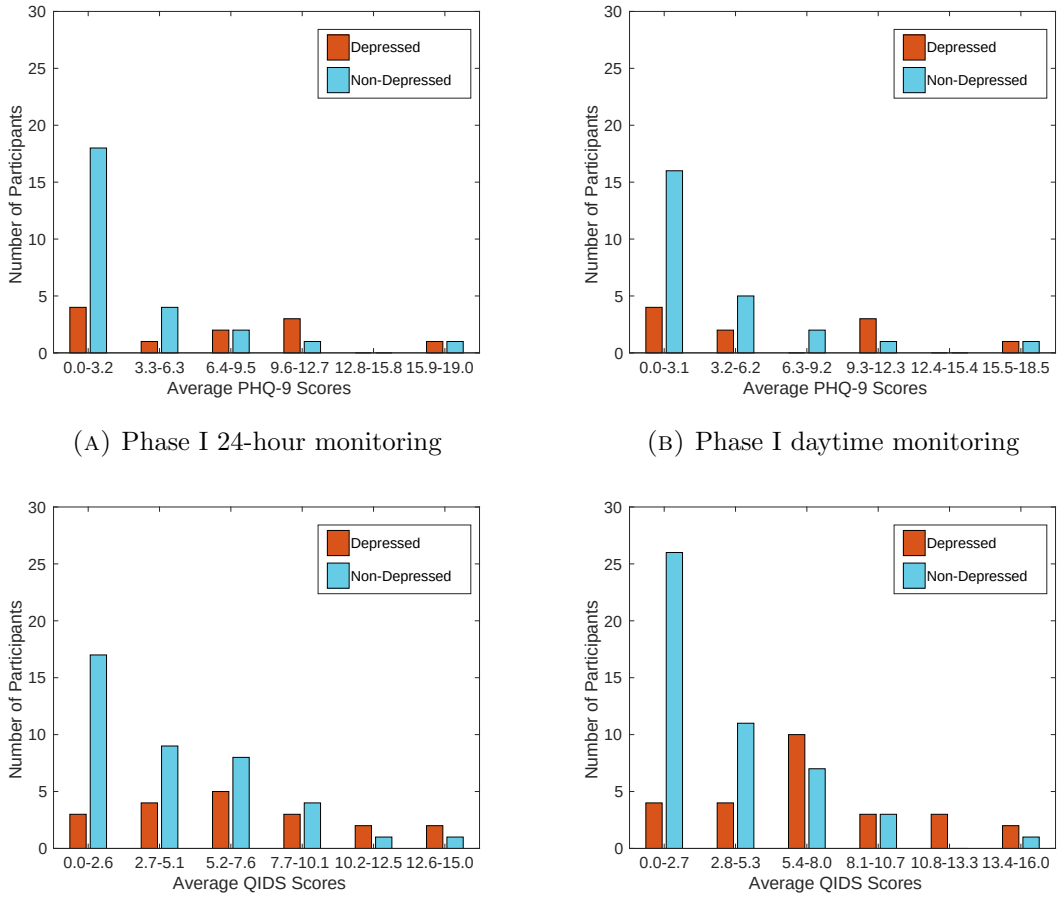


(D) Phase II daytime monitoring

FIGURE 2.5: Histogram of the number of self-report intervals contributed by a participant.

Number of self-report intervals contributed by a user. Figure 2.5(A) plots the histogram of the number of PHQ-9 intervals contributed by a participant in Phase I study 24-hour monitoring. It shows that most of the participants contributed 2-6 PHQ-9 intervals. Figure 2.5(B) plots the results for Phase I daytime monitoring, showing most of the participants contributed 1-5 PHQ-9 intervals. Figures 2.5(C)

and (D) plot the corresponding results for Phase II study, and shows that most of the participants contributed 1-6 QIDS intervals.



(A) Phase I 24-hour monitoring

(B) Phase I daytime monitoring

(C) Phase II 24-hour monitoring

(D) Phase II daytime monitoring

FIGURE 2.6: Histogram of the average self-report scores.

Self-report scores. Since different self-report intervals are included for the analysis for different scenarios, Figure 2.6 plots the histogram of the self-report scores for the different scenarios. In each scenario, for a participant, we plot his/her average self-report score of all the self-report scores considered in that scenario. We see that participants with depression indeed tend to have higher PHQ-9/QIDS scores. We

also see that for the depressed participants, there is a general decreasing trend in self-report scores over time, which is consistent with the fact that all depressed participants were under treatment (they were required to be under treatment to be in the study); for the non-depressed participants, there is no clear trend. The corresponding figures are omitted in the interest of space.

2.5.2 AP Level Features

We extract the following features based on the APs that a participant visited over a given PHQ-9 or QIDS interval. Each AP is considered as a unique location.

Entropy. Entropy measures the variability of time that a participant spends at different APs. Let p_i denote the percentage of time that a participant spends at AP i. The entropy is calculated as

$$\text{Entropy} = -\sum (p_i \log p_i) \tag{2.1}$$

Normalized entropy. Since the number of APs that a participant visited during a PHQ-9 or QIDS interval varies, and entropy increases as the number of APs increases, we also adopt normalized entropy [61], which is invariant to the number of APs and depends solely on the distribution of the visited APs. It is calculated as

$$\text{Entropy}_N = \text{Entropy}/\log N_{\text{loc}} \tag{2.2}$$

where N_{loc} is the number of unique APs that a participant visited during a PHQ-9 or QIDS interval, as to be described below.

Number of unique APs. This feature, denoted as N_{loc}, represents the number of unique APs that a participant visited in a PHQ-9 or QIDS interval.

Time spent at home. We use the approach described in [61] to identify the "Home" AP for a participant as the AP where the participant is most frequently found between 12am to 6am. After that, we calculate the percentage of time when a participant is at the home AP, denoted as *Home*. This feature is only included in the scenario of 24-hour monitoring, which contains nighttime data.

Circadian Movement. We adopt circadian movement [61], referred to as *CMove*, to capture the temporal information of the location data. This feature measures to what extent a participant's sequence of locations followed a 24-hour, or circadian rhythm. To calculate circadian movement, we first use the least-squares spectral analysis, also known as the Lomb-Scargle method [55], to obtain the spectrum of the WiFi association data based on the APs visited. We then calculate the amount of energy that falls into the frequency bins within a 24 ± 0.5 hour period as

$$\text{E} = \sum_{i} psd(f_i) / \ (i_1 - i_2) \tag{2.3}$$

where $i = i_1, i_1 + 1, \ldots, i_2$, and i_1 and i_2 represent the frequency bins corresponding to 24.5 and 23.5 hour periods, respectively, $psd(f_i)$ denotes the power spectral density at each frequency bin f_i. The total circadian movement is then calculated as

$$\text{CMove} = \log(E) \tag{2.4}$$

Number of significant locations visited. This featured, referred to as N_{sig}, is adapted from [13]. Let S denote the top 10 most significant APs visited by a user

(i.e., the 10 APs where a user spent the most time) during the period of study. The number of significant locations in a self-report interval (i.e., PHQ-9 or QIDS interval) is the number of unique APs visited in the interval that are in S.

Routine Index. This feature, referred to as *RIndex* henceforth, is adapted from [13]. It considers a self-report interval (i.e., PHQ-9 or QIDS interval), and quantifies how different the APs visited by a user in a day differs from those visited in another day. Specifically, consider two days d_1 and d_2. Let $\ell_{i1}, \ldots, \ell_{in}$ denote the APs that were visisted in each minute on day i, $i = 1, 2$ (we only consider the set of intervals where there are recorded locations in both days). Then the similarity of these two days is

$$sim(d_1, d_2) = \left(\sum_{j=1}^{n} g(\ell_{1j}, \ell_{2j}) \right) / n$$

where $g(\ell_{1j}, \ell_{2j}) = 1$ if $\ell_{1j} = \ell_{2j}$, and is zero otherwise. We see the value of $sim(d_1, d_2)$ is between 0 and 1, and a larger value represents a higher degree of similarity. Then the routine index of a self-report interval is the average of the similarities of all pairs of days within the interval. It is a value between 0 and 1; higher values indicate that the locations visited over the days are more similar.

2.5.3 Data Analysis

In the following, we first analyze the correlation between the various features and the self-report scores. We then develop regression models to predict the self-report scores, and develop classification models to predict depression status.

TABLE 2.1: AP level analysis: correlation between features and self-report scores.

		All		**Depressed**		**Non-depressed**	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
Phase I 24-hour monitoring	Entropy	-0.36	0.00	-0.40	0.01	-0.24	0.009
	Entropy_N	-0.33	0.00	-0.44	5×10^{-3}	-0.06	0.51
	Home	0.22	6×10^{-3}	0.40	0.01	-0.20	0.03
	N_{loc}	-0.26	9×10^{-4}	-0.22	0.10	-0.36	10^{-4}
	CMove	0.01	0.86	-0.20	0.22	0.12	0.19
	N_{sig}	-0.13	0.12	-0.16	0.36	0.008	0.93
	RIndex	0.32	10^{-4}	0.34	0.03	0.05	0.60
Phase I Daytime monitoring	Entropy	-0.37	10^{-4}	-0.33	0.02	-0.39	3×10^{-4}
	Entropy_N	-0.36	10^{-4}	-0.41	0.02	-0.22	0.04
	N_{loc}	-0.24	0.04	-0.09	0.60	-0.41	10^{-4}
	CMove	-0.19	0.04	-0.23	0.24	-0.11	0.32
	N_{sig}	-0.12	0.21	-0.10	0.60	-0.02	0.84
	RIndex	0.46	0.00	0.37	0.05	0.51	0.00
Phase II 24-hour monitoring	Entropy	-0.05	0.30	0.08	0.40	-0.14	0.09
	Entropy_N	-0.05	0.30	0.17	0.10	-0.16	0.04
	Home	0.13	0.04	-0.13	0.30	0.22	4×10^{-3}
	N_{loc}	0.006	0.90	-0.15	0.20	0.02	0.80
	CMove	-0.18	7×10^{-3}	-0.12	0.35	-0.19	0.01
	N_{sig}	0.20	2×10^{-3}	0.17	0.19	0.17	0.04
	RIndex	-0.08	0.24	-0.27	0.03	0.03	0.73
Phase II Daytime monitoring	Entropy	-0.11	0.10	-0.10	0.30	-0.07	0.30
	Entropy_N	-0.11	0.09	-0.02	0.80	-0.11	0.18
	N_{loc}	-0.03	0.60	-0.12	0.20	0.02	0.81
	CMove	-0.09	0.10	-0.09	0.42	-0.05	0.50
	N_{sig}	0.09	0.10	0.13	0.20	0.02	0.70
	RIndex	0.13	0.04	0.03	0.78	0.16	0.05

Correlation Analysis

We calculated Pearson's correlation coefficients between WiFi meta-data features and self-report scores (PHQ-9 for Phase I study and QIDS for Phase II study). The first half of Table 2.1 presents the correlation results along with p-values (using significance level $\alpha = 0.05$) for Phase I study. The results for both 24-hour and daytime monitoring are shown in the table. Specifically, the results are for three cases: one for all participants, another for depressed participants only, and the third for non-depressed participants only. We observe that the correlation between a feature and the self-report score tends to be higher for depressed participants than that for all participants, and the correlation results for non-depressed participants tend to be the lowest in the three cases (except for the number of unique locations in both 24-hour and daytime monitoring, and routine index in daytime monitoring). This is consistent with the observations in [49], which shows similar results when using data collected directly on smartphones. As speculated in [49], this might be because variation in self-report scores among non-depressed participants may be due to incidental variations in lifestyle rather than psychological changes associated with depression, and hence the correlations between the features and self-report scores are weaker.

For Phase I 24-hour monitoring, we observe that four features, entropy, normalized entropy, the amount of time at home, and routine index, have significant correlation with the self-report (PHQ-9) scores. The significant negative correlation between entropy and self-report scores indicates that participants with relatively high PHQ-9 scores tend to spend more time in a few locations (the same holds for normalized entropy); the positive correlation between time spent at home and PHQ-9 scores suggests that they tend to spend more time at home. These observations are consistent

with existing studies that show depression is associated with social isolation [11, 62]. They are also consistent with earlier studies [61, 25] that use data directly captured on smartphones, indicating that the features obtained from WiFi meta-data provide similar insights into human behavior as those directly obtained from phones. Routine index shows significant positive correlation with self-report scores for depressed participants, maybe because depressed participants tend to be in fewer locations, and tend to spend more time at home. The correlation results under Phase I daytime monitoring are similar as those under 24-hour monitoring.

The second half of Table 2.1 presents the correlation results for Phase II study. We see that the correlation results tend to be much lower compared to those in Phase I. For 24-hour monitoring, only the number of significant locations shows moderate correlation with self-report scores for all participants; and for depressed participants, only routine index shows moderate correlation with self-report scores. For daytime monitoring, none of the features show correlation beyond ± 0.20. The much weaker correlation between the features and the self-report scores in Phase II study may be because the features are obtained from location data during a QIDS interval, which is approximately half of the length of a PHQ-9 interval. The aggregate location features calculated in a short time period may be more subject to noises, and hence show less significant correlation with the self-report scores. On the other hand, as we shall see later on, while individual features in Phase II study do not have significant correlation with self-report scores, they collectively provide reasonably good prediction of the self-report scores and depression status.

Multi-Linear Regression Results

We used the multiple behavioral features to jointly predict self-report scores, and investigated whether they collectively have a stronger correlation with self-report scores. Specifically, we applied both a linear multi-linear regression model, ℓ_2-regularized ϵ-SV (support vector) multivariate regression [23], and a non-linear multi-linear regression model, radial basis function (RBF) ϵ-SV multivariate regression [14], both using the features described above to estimate the self-report scores.

Throughout, we used leave-one-user-out cross validation (i.e., the data of one user was either used for training or testing, but never for both, to avoid overfitting the models since the data of a user over different PHQ-9/QIDS intervals may be correlated) to optimize the model parameters and report the resulting correlation. For ℓ_2-regularized ϵ-SV regression, the parameters to be optimized include the cost parameter C, which is varied over an exponential sequence of values $2^{-10}, 2^{-9}, ..., 2^{10}$, and the margin ϵ, which is varied in $[0, 5]$. For RBF ϵ-SV regression, the parameters to be optimized include cost parameter C, the margin ϵ, and the parameter γ of the radial basis function; the first two parameters are varied in the same manner as those for ℓ_2-regularized ϵ-SV regression, and the last parameter is selected from $2^{-15}, 2^{-9}, ..., 2^{15}$. To assess the performance for each model, we calculated Pearson's correlation after cross validation, which allows us to compare with the results when using single features in Table 2.1.

Table 2.2 summarizes the regression results for four cases, Phase I and Phase II, with 24-hour and daytime monitoring for both cases. We observe that for all the four cases, compared to the linear model, the regressed value from the non-linear model has a much stronger correlation with the ground-truth self-report scores, demonstrating

TABLE 2.2: AP level analysis: multi-feature regression results.

		Phase I		Phase II	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.20	0.01	0.15	0.02
	Multi-feature model (RBF)	0.51	0.00	0.50	0.00
Daytime monitoring	Multi-feature model (linear)	0.17	0.06	0.14	0.04
	Multi-feature model (RBF)	0.64	0.00	0.42	0.00

that the nonlinear model significantly outperforms the linear model. We also observe that the prediction results in Phase I tends to be better than the corresponding results in Phase II, particularly for daytime monitoring (0.64 vs. 0.42 under the non-linear models). This trend is consistent with the significant lower correlation between individual features and self-report scores in Phase II, compared to that in Phase I. On the other hand, for all the four cases with non-linear models, the correlation of the multi-feature regressed value with the self-report scores is significantly larger than that under individual features, indicating that the multiple features are complementary to each other, and combining them significantly improves upon a model that use a single feature.

The correlations of the regressed self-report scores with the ground-truth values as reported above (in both Phases I and II) are comparable or larger than those in [61, 25, 87] (where the correlation range from 0.23 to 0.63), which use the data collected by instrumenting phones. The above results indicate that data collected from the WiFi infrastructure have similar prediction capability as those collected directly from phones.

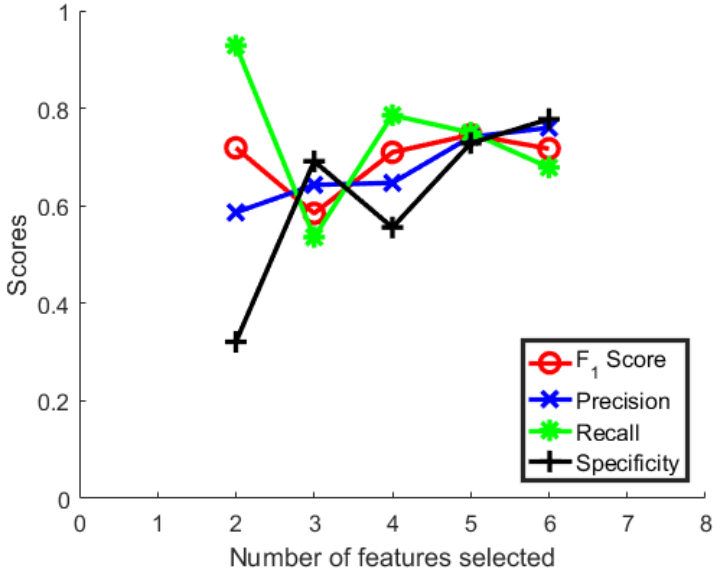


FIGURE 2.7: Illustration of the variation of F_1 score when increasing the number of selected features.

Classification Results

We trained Support Vector Machine (SVM) models with a RBF kernel [14] for classifying whether one is depressed or not, where the assessment from the study clinician is used as the ground truth. Specifically, we considered the depressed class as positive and the non-depressed class as negative, and used leave-one-user-out cross validation (i.e., no data from one user was used in both training and testing to avoid overfitting) procedure to choose parameters for the SVM model. Specifically, the SVM model has two hyper-parameters, the cost parameter C and parameter γ of the radial basis function. We varied the two parameters, C and γ, both from $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and chose the values that gave the best validation F_1 score. The F_1 score, $= 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, can be interpreted as a weighted average of the precision and recall. It ranges from 0 to 1, and the higher,

the better.

The above choice of parameters is performed for a given set of features. To select features, we used SVM recursive feature elimination (SVM-RFE) [33, 57, 86], which is a wrapper-based feature selection algorithm designed for SVM. The goal of SVM-RFE is to find a subset of features out of all the features to maximize the performance of the SVM predictor. For a set of n features (in our context, n is 7 and 6 for 24-hour and daytime monitoring, respectively), we used SVM-RFE for feature selection as follows. For each pair of C and γ values, SVM-RFE provided a ranking of the features, from the most important to the least important. After that, for each feature, we obtained its average ranking across all the combinations of C and γ values, leading to a complete order of the features. Let $\hat{f}_1, \ldots, \hat{f}_n$ represent the n features in descending order of importance. That is, on average, $\hat{f}_1$ is the most important feature, $\hat{f}_2$ is the second most important feature, ..., and $\hat{f}_n$ is the least important feature. We then vary the number of features, k, from 1 to n. For a given k, the features $\hat{f}_1, \ldots, \hat{f}_k$ were used to choose the parameters, C and γ, to maximize F_1 score based on the leave-one-user-out cross validation procedure as described above. Figure 2.7 shows an example (for Phase I study with daytime monitoring) when varying the number of selected features. It plots the F_1 score, along with precision, recall and specificity, as k increases from 2 to 6. We see that $k = 5$ leads to the best F_1 score, and hence the corresponding five features are selected, and the corresponding F_1 score is recorded.

TABLE 2.3: AP level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	RIndex, Entropy	Entropy, Entropy_N, N_{loc}, CMove, N_{sig}
Phase II	CMove, Entropy_N, RIndex, N_{loc}	CMove, Entropy, N_{loc}, Entropy_N, N_{sig}, RIndex

TABLE 2.4: AP level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.66	0.63	0.70	0.58	0.74	0.74	0.75	0.72
PHQ-9 (Phase I)	0.68	0.61	0.75	0.53	0.72	0.67	0.78	0.60
Features (Phase II)	0.78	0.86	0.72	0.85	0.79	0.73	0.88	0.53
QIDS (Phase II)	0.72	0.68	0.76	0.70	0.85	0.84	0.86	0.85

We next present the classification results for the value k that provided the highest F_1 score in the four scenarios: Phase I and Phase II, with 24-hour and daytime monitoring for both studies. Table 2.3 lists the top k features. We observe that for Phase I study, entropy is a selected feature for both 24-hour and daytime monitoring, consistent with its significant correlation with the self-report scores (see the first half of Table 2.1). For Phase II study, while the correlation between a single feature and the self-report score is generally weak, the features that are selected do have relatively high correlation in certain cases (see the second half of Table 2.1).

Table 2.4 shows the F_1 score along with three other performance metrics (precision, recall and specificity) for the four scenarios. The F_1 score is 0.66-0.79. Maybe surprisingly, the results for Phase II study is comparable (even slightly better) than those for Phase I study, despite the weaker correlation between the features and the self-report scores. For comparison, Table 2.4 also lists the classification results when using self-report scores (i.e., PHQ-9 scores for Phase I and QIDS scores for Phase II), where we chose an optimal threshold for classification. We observe that the classification results when using the features from WiFi meta-data are comparable to those when using self-reports (as we shall see in Section 2.6, the features at the building level can lead to even better classification results than using self-reports). Given that WiFi meta-data are collected automatically, which does not require users to fill in the questionnaires or direct data collection on the phones, our results demonstrate

that using WiFi meta-data can be a promising light-weight and low-cost approach for automatic depression screening.

Overall, the classification results are comparable to those in [61, 25, 87], which use data collected directly from smartphones, indicating that data collected from the WiFi infrastructure can lead to similar classification accuracy. The results for one setting, Phase I 24-hour monitoring, are worse than other settings; as we shall see in Section 2.6.3, it is significantly improved when using building based features.

2.6 Building Level Analysis

In this section, we present analysis results on the building level. Specifically, if a WiFi association record indicates that a user is associated with an AP a from time t to t', then we map the AP to the corresponding building b, and regard that the user is in building b from t to t'. In the following, we first present our data preprocessing procedure, and then describe feature extraction and analysis results. As mentioned earlier, the reason for using building based features is that intuitively they may represent the location more meaningfully (when a user is associated with different APs in the same building, he/she is essentially at the same location semantically).

2.6.1 Data Preprocessing

We preprocess the data following a similar methodology as that in Section 2.5.1. For the data collected in Phase I study, we consider PHQ-9 intervals. For each PHQ-9 interval, we only include the buildings where a participant spent at least one hour over the PHQ-9 interval (to avoid including locations that a participant simply passed

by). For 24-hour monitoring, the results below are for the case when we include a PHQ-9 interval into analysis if it has at least 14 days of data; for daytime monitoring, the threshold is 13 days (we use a lower threshold to cover more users and PHQ-9 intervals). For 24-hour monitoring, we obtained a total of 146 PHQ-9 intervals. Out of these, 36 belonged to depressed participants and 110 belonged to the non-depressed participants. A total of 37 users are found in this dataset, with 11 as depressed and 26 as non-depressed. For daytime setting, we extracted a total of 155 PHQ-9 intervals. Out of these, 37 PHQ-9 intervals belonged to depressed participants and 118 PHQ-9 intervals belonged to non-depressed participants; we found 43 users in this setting, with 13 as depressed and 30 as non-depressed.

For the data collected in Phase II study, we consider QIDS intervals. For each QIDS interval, we only include the buildings where a participant spent at least 30 minutes over the QIDS interval. For both 24-hour and daytime monitoring, QIDS intervals with at least 7 days of data are included for the analysis. For 24-hour monitoring, we extracted a total of 216 QIDS intervals, with 64 QIDS intervals belonging to depressed participants and 152 belonging to non-depressed participants. These QIDS intervals are obtained from a total of 59 users, with 19 as depressed and 40 as non-depressed. In daytime monitoring, we obtained 212 QIDS samples, with 68 belonging to depressed participants and 144 belonging to non-depressed participants. There are 74 users, with 26 as depressed and 48 as non-depressed. Overall, the dataset for Phase II study is larger than that of the Phase I study.

The time coverage (i.e., the percentage of time with WiFi association data) is similar as that for AP level analysis for all the above four scenarios. The number of intervals contributed by a participant is also similar as that in AP level analysis. The figures are omitted for clarity.

2.6.2 Feature Extraction

Based on the buildings that a participant visited over a given PHQ-9 or QIDS interval, we extracted the following features: entropy, normalized entropy, the number of unique buildings visited, the amount of time that a user spent in the "home" building (only for 24-hour monitoring), circadian movement, the number of significant buildings visited, and routine index. These features are defined as those in Section 2.5.2, except that they use building based locations instead of AP based locations. As noted earlier, semantically, building based location is more meaningful. However, it requires additional knowledge on which building an AP belongs to.

2.6.3 Data Analysis

The data analysis below proceeds in the same order as that in Section 2.5.3: we first report the correlation between the various features and the self-report scores, followed by the multi-feature regression results for predicting the self-report scores and classification results for predict depression status.

Correlation Analysis

We computed the correlation results using Pearson's correlation coefficients between the various building-level features and self-report scores. The first half of Table 2.5 presents the correlation results along with p-values (using significance level $\alpha = 0.05$) for Phase I study. Again, we see that the correlations tend to be stronger for depressed participants compared to those for all participants, and non-depressed participants. For 24-hour monitoring, all the seven features show significant correlation with self-report scores; the correlations are particularly significant for depressed participants.

TABLE 2.5: Building level analysis: correlation between features and self-report scores.

		All		Depressed		Non-depressed	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
Phase I 24-hour monitoring	Entropy	-0.28	4×10^{-4}	-0.50	10^{-3}	-0.31	9×10^{-4}
	Entropy_N	-0.28	5×10^{-4}	-0.51	10^{-3}	-0.26	5×10^{-3}
	Home	0.28	6×10^{-4}	0.51	10^{-3}	0.26	6×10^{-3}
	N_{loc}	-0.21	7×10^{-3}	-0.34	0.04	-0.31	10^{-3}
	CMove	-0.21	0.01	-0.47	3×10^{-3}	-0.12	0.20
	N_{sig}	-0.26	10^{-3}	-0.32	0.05	-0.30	10^{-3}
	RIndex	0.32	10^{-4}	0.41	0.01	0.35	10^{-4}
Phase I Daytime monitoring	Entropy	-0.27	7×10^{-4}	-0.38	0.01	-0.31	7×10^{-4}
	Entropy_N	-0.29	2×10^{-4}	-0.40	0.01	-0.27	2×10^{-3}
	N_{loc}	-0.13	8×10^{-3}	-0.12	0.40	-0.26	4×10^{-3}
	CMove	-0.20	0.01	-0.13	0.40	-0.22	0.01
	N_{sig}	-0.13	0.11	-0.07	0.68	-0.26	3×10^{-3}
	RIndex	0.38	0.00	0.36	0.02	0.47	0.00
Phase II 24-hour monitoring	Entropy	-0.17	8×10^{-3}	-0.29	0.01	-0.08	0.31
	Entropy_N	-0.20	2×10^{-3}	-0.31	0.01	-0.1	0.22
	Home	0.24	3×10^{-4}	0.46	10^{-4}	0.13	0.11
	N_{loc}	-0.07	0.20	-0.16	0.10	-0.04	0.62
	CMove	-0.008	0.90	-0.15	0.20	-0.02	0.81
	N_{sig}	0.13	0.06	-0.05	0.60	0.21	0.00
	RIndex	0.11	0.12	0.32	0.01	0.01	0.82
Phase II Daytime monitoring	Entropy	-0.17	0.01	-0.24	0.04	-0.04	0.58
	Entropy_N	-0.20	2×10^{-3}	-0.30	0.01	-0.05	0.49
	N_{loc}	-0.07	0.20	-0.09	0.40	-0.03	0.68
	CMove	-0.04	0.50	-0.33	6×10^{-3}	0.03	0.74
	N_{sig}	-0.02	0.70	-0.12	0.30	0.05	0.53
	RIndex	0.23	0.00	0.30	0.01	0.09	0.28

The sign of the correlation is consistent with the observations [11, 62] that the participants with higher self-report scores tend to spend time in a few places and spend more time at home. For daytime monitoring, we observe significant correlation for entropy, normalized entropy, and routine index, with the signs of the correlations consistent with those of 24-hour monitoring.

The second half of Table 2.5 presents the correlation results for Phase II study. Unlike what we observed for AP level analysis (Section 2.5.3), we observe that several features (entropy, normalized entropy, the amount of time spent at home, and routine index) show significant correlation in various cases. On the other hand, consistent with AP level analysis, the correlations in Phase II are still generally lower than the corresponding values in Phase II, which may be due to the different self-report instruments that were used in these two phases, and particularly, different lengths of the self-report intervals.

For both Phase I and II studies, the above results show that features extracted at the building level show more significant correlation with self-report scores than that at the AP level, consistent with the intuition that buildings represent the visited locations more meaningfully than APs.

Multi-Linear Regression Results

We used multi-linear regression to predict self-report scores using building based features. The approach is similar to what we have described in Section 2.5.3. Again we used leave-one-user-out cross validation. The only difference is that we have now considered the building level features, instead of AP level features. Table 2.6 summarizes the regression results. Similar to what we observed in Section 2.5.3, the

results from the non-linear regression models are significantly better than those from the linear models. For the non-linear models, the r-values range from 0.30 to 0.46 across the four scenarios (i.e., Phase I and Phase II studies with 24-hour and daytime monitoring in both cases), all with small p-values. In addition, for each scenario, the r-values obtained from the ℓ_2-regularized non-linear models are better than the corresponding r-values obtained using individual features (see Table 2.5). Again, the results for Phase I are better than those for Phase II, consistent with the stronger correlation for individual features observed in Section 2.6.3.

The regression results under the linear models are similar as those for the AP level (see Section 2.5.3). Somewhat surprisingly, the regression results for the non-linear models are worse than those for the AP level, despite the stronger correlation between the individual features and the self-report scores at the building level, which might be due to the relative small sample size (particularly the small number of depressed participants). On the other hand, the r values are still comparable or higher than those in [61, 25, 87], which are obtained using data collected directly from phones. In addition, as we shall see next, the building level features lead to better classification results than AP level features.

TABLE 2.6: Building level analysis: multi-feature regression results.

		Phase I		**Phase II**	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.22	0.00	0.13	0.05
	Multi-feature model (RBF)	0.46	0.00	0.37	0.00
Daytime monitoring	Multi-feature model (linear)	0.20	0.01	0.10	0.10
	Multi-feature model (RBF)	0.46	0.00	0.30	0.00

Classification Results

TABLE 2.7: Building level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	CMove, N_{sig}, N_{loc}, RIndex, Entropy	N_{loc}, Entropy$_N$, CMove, N_{sig}, Entropy, RIndex
Phase II	RIndex, N_{loc}, Entropy	Entropy$_N$, Entropy, N_{loc}, RIndex, CMove, N_{sig}

The classification approach is similar to what we have described in Section 2.5.3, except for that the features are based on buildings instead of APs. We again used leave-one-user-out cross validation to determine the two hyper-parameters, and used SVM-RFE to select features. Table 2.7 lists the top k features selected by SVM-RFE for various scenarios. For daytime monitoring, in both Phase I and II studies, all the six features have been selected, which provided the best F_1 score. For 24-hour monitoring, a subset of features are selected, and the number of unique buildings, entropy and routine index have been selected for both Phase I and II studies. Table 2.8 summarizes the classification results. The F_1 score is 0.73-0.84 in various scenarios. For comparison, we again list the classification results when using self-report scores. We see that in two cases (24-hour monitoring, Phase I and II studies), the classification results obtained using the features are substantially better than those obtained using the self-report scores; for the other two cases, the classification results obtained using these two approaches are similar. The above results again confirm that automatic classification using the WiFi association based features is a promising way for automatic depression screening.

Compared to the classification results for AP level analysis (Section 2.5.3), the results for the building level analysis are substantially better for one scenario (Phase I 24-hour monitoring); the results for other cases are comparable. The above results

indicate that the building level features are probably more meaningful in representing people's behaviors for classification tasks. We further see from Table 2.8 that the classification results under 24-hour monitoring tend to be better than daytime monitoring, which is perhaps not surprising since 24-hour monitoring uses the data from both night and day, while only partial data (8am-6pm) is used in daytime monitoring.

TABLE 2.8: Building level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.84	0.90	0.77	0.90	0.75	0.67	0.80	0.62
PHQ-9 (Phase I)	0.68	0.55	0.88	0.53	0.70	0.63	0.78	0.57
Features (Phase II)	0.79	0.84	0.75	0.82	0.73	0.73	0.73	0.63
QIDS (Phase II)	0.67	0.57	0.81	0.50	0.85	0.86	0.85	0.87

2.7 Enhanced Building Level Analysis

In this section, we enhance the building level analysis in the previous section by considering several additional building level features, which are related to the categories of the buildings. Our goal is to investigate whether including these additional features can further improve the prediction results.

2.7.1 Additional Building Level Features

All the additional features are based on the categories of the buildings. Specifically, we broadly classified the campus buildings based on their main purposes as entertainment, sports, class, library, and others. For each category of buildings, we extract three types of features, detailed as follows.

Number of Entertainment, Sports and Class buildings visited. The campus has multiple entertainment, sports and class buildings. For each category of buildings, we calculated the number of unique buildings visited by a participant in a given PHQ-9 or QIDS interval. These features are denoted as N_{entr}, N_{sports}, and N_{class}, respectively.

Average duration spent in Entertainment, Sports, Library and Class buildings. These features represent the average duration that a participant spent in each category of buildings over a PHQ-9 or QIDS interval. They are denoted as D_{entr}, D_{sports}, D_{library}, and D_{class}, respectively.

Number of days visiting Entertainment, Sports, Library and Class buildings. These features represent the number of days that a participant visited a specific category of buildings over a PHQ-9 or QIDS interval. They are denoted as Day_{entr}, Day_{sports}, Day_{library}, and Day_{class}, respectively.

Table 2.9 presents the correlation of these additional features with self-report scores for Phase I data. For 24-hour monitoring, we observe one feature, the number of days visiting Entertainment buildings, has significant correlation with the self-report scores for both all and depressed participants. One feature (the number of class buildings visited) shows significant correlation for all participants, but not for the depressed participants; several other features (the duration in library, the number of days visiting sports buildings and library) show significant correlation for the depressed participants, but not for all participants. For daytime monitoring, some features show significant correlation for all participants, some features show significant correlation for the depressed participants, while no feature shows significant correlation for both all and depressed participants.

TABLE 2.9: Enhanced building level analysis: correlation between building-category features and self-report scores for Phase I study.

		All		**Depressed**		**Non-depressed**	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	N_{entr}	-0.07	0.30	-0.04	0.80	-0.14	0.12
	N_{sports}	-0.06	0.40	-0.21	0.20	-0.14	0.14
	N_{class}	-0.31	10^{-4}	0.11	0.50	-0.37	10^{-4}
	D_{entr}	0.03	0.70	-0.18	0.20	0.10	0.26
	D_{sports}	-0.05	0.50	-0.22	0.10	-0.08	0.37
	D_{library}	0.05	0.50	-0.34	0.04	0.34	2×10^{-4}
	D_{class}	-0.12	0.10	0.11	0.06	-0.26	0.004
	Day_{entr}	-0.21	9×10^{-3}	-0.29	0.07	-0.28	0.002
	Day_{sports}	-0.08	0.20	-0.32	0.05	-0.12	0.20
	Day_{library}	-0.01	0.80	-0.33	0.04	0.19	0.04
	Day_{class}	-0.36	10^{-4}	-0.30	0.06	-0.42	0.00
Daytime monitoring	N_{entr}	-0.09	0.02	0.01	0.90	-0.23	0.01
	N_{sports}	-0.07	0.30	-0.22	0.10	-0.04	0.63
	N_{class}	-0.22	5×10^{-3}	0.13	0.40	-0.31	5×10^{-4}
	D_{entr}	-0.19	0.01	-0.17	0.20	-0.24	0.007
	D_{sports}	-0.04	0.50	-0.16	0.30	-0.04	0.64
	D_{library}	0.09	0.20	-0.35	0.03	0.40	0.00
	D_{class}	-0.09	0.20	0.30	0.06	-0.27	0.002
	Day_{entr}	-0.15	6×10^{-3}	-0.17	0.30	-0.25	0.004
	Day_{sports}	-0.10	0.10	-0.30	0.07	-0.07	0.44
	Day_{library}	0.009	0.90	-0.21	0.20	0.15	0.08
	Day_{class}	-0.28	3×10^{-4}	-0.18	0.20	-0.34	10^{-4}

TABLE 2.10: Enhanced building level analysis: correlation between features and self-reports for Phase II study.

		All		**Depressed**		**Non-depressed**	
	Features	**r-value**	**p-value**	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	N_{entr}	-0.10	0.10	-0.21	0.09	-0.07	0.33
	N_{sports}	-0.01	0.70	-0.04	0.70	0.02	0.75
	N_{class}	-0.03	0.50	-0.04	0.70	0.09	0.26
	D_{entr}	-0.11	0.10	-0.31	0.01	-0.01	0.81
	D_{sports}	-0.01	0.70	-0.07	0.50	0.04	0.58
	D_{library}	-0.12	0.60	-0.08	0.40	-0.15	0.05
	D_{class}	0.14	0.03	0.06	0.60	0.17	0.02
	Day_{entr}	0.003	0.90	-0.12	0.30	0.01	0.83
	Day_{sports}	-0.0004	0.90	0.05	0.60	-0.02	0.72
	Day_{library}	-0.17	9×10^{-3}	0.10	0.30	-0.22	4×10^{-3}
	Day_{class}	-0.01	0.80	-0.02	0.80	0.09	0.25
Daytime monitoring	N_{entr}	-0.13	0.04	-0.15	0.20	-0.17	0.03
	N_{sports}	-0.03	0.60	-0.05	0.60	0.006	0.94
	N_{class}	-0.07	0.20	-0.10	0.30	0.07	0.35
	D_{entr}	-0.11	0.08	-0.22	0.06	-0.07	0.37
	D_{sports}	-0.05	0.40	-0.12	0.30	-0.003	0.96
	D_{library}	-0.12	0.07	-0.04	0.70	-0.12	0.12
	D_{class}	0.09	0.10	0.03	0.70	0.14	0.08
	Day_{entr}	-0.07	0.30	-0.14	0.20	-0.09	0.25
	Day_{sports}	-0.04	0.50	0.02	0.80	-0.04	0.57
	Day_{library}	-0.13	0.04	0.01	0.80	-0.14	0.07
	Day_{class}	-0.05	0.30	-0.04	0.70	0.07	0.39

Table 2.10 presents the correlation results for Phase II study. We only observe a few cases (the duration in entertainment buildings for depressed participants for both 24-hour and daytime monitoring) with significant correlation; the rest of the cases have low correlation.

Again, the differences in the correlation results for Phases I and II may be caused by the different self-report instruments, and particularly different lengths of the self-report intervals in these two phases. Overall, the correlation of the various features with the self-report scores is not very strong. On the other hand, as we shall see, they are still helpful in improving classification results.

2.7.2 Multi-Linear Regression Results

The multi-linear regression approach is similar to what we have described earlier (Sections 2.5.3 and 2.6.3). Again we used leave-one-user-out cross validation. The only difference is that we have now considered the building level features (Section 2.6.2) together with the various building category features (Section 2.7.1). Table 2.11 summarizes the regression results. Similar to what we have observed earlier, the results from the non-linear regression models are better than those from the linear models; and multi-feature regression improves upon single-feature models. Compared to the building level analysis that does not include building category features (Section 2.6.3), we see that the performance becomes slightly worse, indicating that the additional building category features have not helped in improving the regression results. On the other hand, the range of the r values is still comparable to the range obtained by using data directly from the phones [61, 25, 87].

TABLE 2.11: Enhanced building level analysis: multi-feature regression results.

		Phase I		Phase II	
	Model	**r-value**	**p-value**	**r-value**	**p-value**
24-hour monitoring	Multi-feature model (linear)	0.19	0.02	0.14	0.04
	Multi-feature model (RBF)	0.43	0.00	0.36	0.00
Daytime monitoring	Multi-feature model (linear)	0.23	0.00	0.09	0.10
	Multi-feature model (RBF)	0.32	0.00	0.26	0.00

2.7.3 Classification Results

The classification procedure is as that in Section 2.6.3, except that both aggregate building level features and building category features are used for classification. Table 2.12 lists the top k features selected by SVM-RFE for various scenarios. We see that, despite the large number of features, only up to five features are selected in the various scenarios. In addition, a mixture of aggregate building level features and building category features are selected for each scenario. One feature, the number of significant buildings visited (N_{sig}), is selected as one of the top features for all scenarios. Routine index is also selected consistently. For building category features, certain features related to library and sports also tend to be selected.

Table 2.13 summarizes the classification results, showing that the F_1 score ranges from 0.72-0.85 in various scenarios. Compared to the results when not including building category features (see Section 2.6.3), the results for one scenario (Phase II 24-hour monitoring) are improved (the F_1 score is improved from 0.79 to 0.85), and the results for other scenarios remain similar. The above results indicate that adding building category features can further improve the classification performance.

TABLE 2.12: Enhanced building level analysis: top features selected by SVM-RFE.

	24-hour monitoring	**Daytime monitoring**
Phase I	RIndex, N_{sig}, CMove, $D_{library}$, $Day_{library}$	N_{sig}, $Day_{library}$
Phase II	RIndex, N_{sig}	RIndex, D_{sports}, $Day_{library}$, N_{sig}

TABLE 2.13: Enhanced building level analysis: classification results.

	24-hour monitoring				**Daytime monitoring**			
	F_1 Score	Precision	Recall	Specificity	F_1 Score	Precision	Recall	Specificity
Features (Phase I)	0.83	0.78	0.89	0.75	0.74	0.71	0.78	0.69
PHQ-9 (Phase I)	0.68	0.55	0.88	0.53	0.70	0.63	0.78	0.57
Features (Phase II)	0.85	0.88	0.82	0.86	0.72	0.68	0.76	0.50
QIDS (Phase II)	0.67	0.57	0.81	0.50	0.85	0.86	0.85	0.87

2.8 Conclusion and Future Work

In this part of the dissertation, we have investigated using meta-data passively collected from WiFi infrastructure for automatic depression screening. We have extracted various features at both the AP and building levels, and investigated their correlations with self-report scores. In addition, we have constructed a family of machine learning models for predicting self-report scores and depression status. Our analysis over two datasets demonstrated that this approach can lead to accurate depression prediction. The prediction results are comparable to those obtained using data collected by instrumenting individual phones. Our study was conducted in a university setting, considering college students, a specific demographic group that has heightened risk of mental health issues including depression [75]. Future directions

include exploring the approach in other university campuses, and in other settings (e.g., company, military base) with different demographic groups.

Acknowledgments

This work was partially supported by the National Science Foundation (NSF) grant IIS-1407205. Jinbo Bi was also supported by the National Institutes of Health (NIH) grants R01DA037349 and K02DA043063, and NSF grants CCF-1514357 and IIS-1718738. The authors thank University of Connecticut Information Technology Services for providing us the WiFi infrastructure meta-data.

Chapter 3

Predicting Depressive Symptoms using Smartphone Data

3.1 Introduction

Depression is a common yet very serious health problem. It impacts a person physically, emotionally as well as socially, leading to higher medical costs, exacerbated medical conditions, and higher mortality [67, 40, 20]. Depression symptoms manifest in many aspects of daily life, including appetite, interests, energy level, mood, psychomotor behavior, sleep, and even suicidal intent. Currently, survey instruments, such as Patient Health Questionnaire-9 (PHQ-9) [44] and Quick Inventory of Depressive Symptomatology (QIDS) [60], are commonly used to detect depression and keep track of the development of the symptoms. Users need to fill in such questionnaires on a regular basis (e.g., biweekly for PHQ-9 and weekly for QIDS), which is burdensome and difficult to execute on a continuous basis.

The emergence of smartphones as a pervasive computing platform presents a tremendous opportunity of using smartphone data to automatically detect depression, as evidenced by existing studies (e.g., [81, 13, 61, 25]). These studies have demonstrated that sensing data (e.g., location, activity, phone usage) collected passively from smartphones can be used for effective depression screening, since the sensing data provides insights into various behavioral features that are highly correlated with depression. A recent study [83] explored an alternative approach that does not require direct data capture on a user's phone; instead, it leverages metadata collected from an institution's WiFi network (e.g., the campus WiFi network of a university or company). The rationale is that, when a user associates her phone with an access point (AP) in an institution's WiFi network for Internet access, the location of the AP can be used to approximate the location of the phone and hence the user (a phone needs to be close to the AP for the association). Therefore, the AP association records from an institution's WiFi network can be used to locate the users dynamically over time. While the above two approaches differ in the data sources that are being used, they share the common idea that high-level human behaviorial features extracted from smartphone data, whether collected on the phones or from a WiFi infrastructure, can be used to train machine learning models beforehand, and then used for automatic depression screening.

Existing studies focus on binary classification using smartphone data, i.e., classifying whether one is depressed or not. In this part of the dissertation, we make a significant step forward in that we predict individual depressive symptoms, including all major aspects covered by PHQ-9 and QIDS. The predicted individual symptoms provides a detailed picture on one's current depression conditions (both behavioral and cognitive), which can be tremendously helpful for both the users and clinicians.

This approach involves no efforts from the users, and can provide an objective assessment that does not suffer from recall bias. On the other hand, predicting the status of individual depressive symptoms is at a much finer granularity than predicting the overall status of being depressed or not, and hence is more challenging. In addition, smartphone data are primarily behavioral data (e.g., location, activity), while the depressive symptoms can be cognitive in nature (e.g., interests, self-criticism, feeling depressed). It is not clear whether behavioral data can be used to predict cognitive characteristics accurately.

To explore the feasibility of using smartphone data to predict individual depressive symptoms, we have analyzed two categories of data collected from 182 college students in a two-phase study. The first category is smartphone sensing data, collected directly on smartphones (by running an app on the phones), and the second category is meta-data collected from a university campus WiFi network. Both categories include the data collected in two phases, accompanied by PHQ-9 and QIDS questionnaires, respectively, which are reported by the participants. We explore using machine learning techniques to predict the presence or absence of each depression symptom using features extracted from the two categories of smartphone data.

Our study makes the following main contributions.

- We find that sensing data collected directly on smartphones can predict a rich set of depressive symptoms accurately, including both behavioral (appetite, energy, sleep, psychomotor) and cognitive symptoms (interests, self-criticism, feeling depressed, concentration). The predicted F_1 scores can be as high as 0.83, comparable to the F_1 scores obtained for predicting the overall depression status [25, 87, 49, 89]. In addition, we observe stronger prediction results for

depressed participants compared to non-depressed participants.

- We find that meta-data collected from an institution's WiFi infrastructure can also predict a variety of depressive symptoms accurately. Specifically, we explore 24-hour monitoring (for the users who spend time during both night and day on campus, e.g., those who live on campus), and daytime monitoring where only the daytime information (8am-6pm) is available (e.g., for those who are only on campus during daytime). We find that even daytime information is sufficient to provide accurate prediction for a set of depressive symptoms. Our results demonstrate that the meta-data collected from an institution's WiFi infrastructure can be used to keep track of the wellness of a large population at very little cost.

- We further explore predicting finer-level depressive symptoms, e.g., increased or decreased appetite/weight, feeling restless or slowed down, and sleep disturbance (time taken falling asleep, sleep during night, sleeping too much, and waking up too early). Our results demonstrate that even finer-level depressive symptoms (particularly sleep related) can be predicted accurately using smartphone data, with predicted F_1 scores up to 0.86.

The rest of the chapter is organized as follows. Section 3.2 briefly describes the background and our high-level approach. Section 3.3 describes the data collection methodology. Sections 3.4 and 3.5 report our analysis methodology and prediction results of individual depressive symptoms using smartphone sensing data and WiFi infrastructure meta-data, respectively. Section 3.6 presents the results on finer-level depressive symptoms. Section 4.2 briefly describes related work. Last, Section 4.7 concludes the chapter and presents future work.

3.2 Background and High-level Approach

In this section, we briefly describe depressive symptoms, particularly the symptoms in two widely used questionnaires, PHQ-9 and QIDS, which are used in this study. We then describe our high-level approach of using smartphone data collected directly from smartphones, or meta-data collected from an institutions's WiFi infrastructure, to predict depressive symptoms.

3.2.1 Depressive Symptoms

Depressive symptoms manifest in multiple aspects. We use two types of questionnaires, PHQ-9 and QIDS, during the two phases of our study (see Section 3.3). Both questionnaires are widely used in clinical settings for detecting depression and keeping track of the depression symptoms over time. PHQ-9 contains 9 questions, asking about the symptoms in the past two weeks, and hence needs to be filled in by a user every two weeks. QIDS is more comprehensive than PHQ-9. It contains 16 questions, asking about the symptoms in the past week, and hence needs to be filled in every week. For both questionnaires, the questions are on nine broad aspects, including (1) appetite/weight, (2) interests, (3) energy/fatigue, (4) concentration, (5) psychomotor agitation/retardation, (6) self-criticism, (7) feeling sad/depressed, (8) sleep disturbance, and (9) suicidal ideation. The score of each question ranges from 0 to 3, corresponding to none, slight, moderate and severe symptoms, respectively. The total score is the sum of the scores of the individual questions, and hence the minimum score is 0 and the maximum score is 27. For certain symptoms, QIDS asks multiple questions (instead of a single question as in PHQ-9), and the maximum score of the responses to the multiple questions is used when calculating the total score.

In the following, we briefly describe the nine questions in PHQ-9, and describe the finer-level questions in QIDS when applicable.

- **Appetite level.** This question asks about poor appetite or overeating. In QIDS, this question is expanded into four sub-questions: (1) increased appetite, (2) decreased appetite, (3) increased weight, and (4) decreased weight, where (1) and (2) are mutual exclusive (one can only choose one to answer), and (3) and (4) are mutual exclusive.

- **Interest level.** This question checks if there is little interest in other people or activities.

- **Energy/fatigue level.** This question asks whether one has lower energy in doing day-to-day activities.

- **Concentration level.** This evaluates whether one has trouble concentrating or making decisions.

- **Psychomotor agitation/retardation.** This question checks if one is feeling more slowed down or restless than usual. In QIDS, this question is divided into two sub-questions: (1) feeling slowed down, and (2) feeling restless.

- **Self-criticism.** This question evaluates whether one feels bad about herself or that she is letting her family down.

- **Feeling sad/depressed.** This question records whether one is feeling down, depressed, sad or hopeless.

- **Sleep disturbance.** This question checks if one is having trouble sleeping. In QIDS, this question is expanded to four sub-questions: (1) time taken falling

asleep, (2) sleep during night, (3) waking up too early, and (4) sleeping too much.

- **Suicidal ideation.** This question checks if one has any suicidal intent.

3.2.2 High-level Approach

We predict the individual symptoms described above using machine learning models. Specifically, we first collect data and ground truth to train machine learning models. After that, we use testing data to evaluate the prediction accuracy of the models. We consider two scenarios of data collection.

- **Using sensing data collected on smartphones.** In this scenario, the sensing data was passively collected from smartphones, through an app that runs in the background on the phones. A wide variety of sensing data (e.g., location, activity, phone usage) can be collected. We primarily focus on location data in this work. The data was collected over 24 hours each day.

- **Using meta-data collected from institution infrastructure.** In this scenario, meta-data was passively collected from a wireless infrastructure (e.g., campus WiFi network in a university). Specifically, we use WiFi association data to provide information on user locations over time (since a phone needs to be close to an AP for the association, the location of the AP can approximate the location of the phone/user). We further consider two cases in this scenario: (1) using data collected over 24 hours each day, and (2) only using data collected during daytime (8am-6pm). The first case is applicable when a user spends significant amount of time during both night and day on campus (e.g., a

student living on campus), while the second is applicable when a user comes to a campus for work/study during the daytime, and spends the rest of the time off campus. Clearly, 24-hour data provides more insights into a user's behavior than daytime data. We also explore the daytime case since it is common in practice, and it is interesting to explore whether daytime location information alone already provides substantial insights into depression symptoms.

Advantages of our approach. Our approach of using passively collected data to automatically predict individual depressive symptoms eliminates the need for users to manually fill in their depressive symptoms, and hence provides a convenient mechanism for continuous monitoring. In addition, the prediction can provide objective assessment, which does not suffer from recall bias. The second scenario described above can be used for depressive symptom assessment on a large scale. For example, it can be used to estimate the percentage of students feeling depressed in a university. When a university carries out activities to improve the mental health of the students (e.g., hold events to raise the awareness of mental health or advocate best practices to improve mental health), it can be further used to assess the effectiveness of these activities (e.g., by comparing the percentage of students feeling depressed before and after the activities).

Deployment issues. The focus of this study is to explore the feasibility of using data collected in the above two scenarios in predicting depression symptoms. Clearly, user privacy and responsible usage of the data need to be considered carefully in any system that uses our approach, which are beyond the scope of this work; a brief discussion of the pros and cons of using these two types of data and deployment issues is in [83].

3.3 Data Collection

We collected data from a two-phase study at the University of Connecticut. Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017. The participants were full-time students of the university, aged 18-25. We recruited 79 participants in Phase I study (73.9% female and 26.1% male; 62.3% white, 24.6% Asian, 5.8% African American, 5.8% with more than one race, and 1.5% being other or unknown), and recruited 103 participants in Phase II study (76.7% female and 23.3% male; 58.3% white, 25.2% Asian, 3.9% African American, 7.8% with more than one race, and 4.9% being other or unknown).

All participants met with our study clinician for informed consent and initial screening before being enrolled in the study. Based on the clinician assessment, in Phase I study, 19 and 60 participants were classified as depressed and non-depressed, respectively; in Phase II study, the corresponding numbers are 39 and 64. In both phases, we intended to recruit the same number of depressed and non-depressed participants, and were not able to recruit as many depressed participants as intended.

A subset of the data has been used in our prior works [25, 87, 49, 89]. None of them explores predicting individual depressive symptoms as in this study. We next briefly describe four types of data that are used in this study: smartphone sensing data, meta-data from campus WiFi infrastruture, questionnaire responses, and clinician assessment. For user privacy, the identities of the participants were removed and were annotated with random IDs.

3.3.1 Smartphone Sensing Data

The sensing data was collected using *LifeRhythm* [25], an app that we developed for Android and iPhone, the two predominant smartphone platforms. The app runs in the background, passively collecting sensing data with no need of user interaction. The Android version of the app was developed based on an existing publicly available library, Emotion Sense library [45]; for iPhone, the app was developed using Swift from scratch. While a variety of sensing data is collected by the app, we focus on location data in this work. Specifically, we collected two types of location data on the phones: GPS locations and WiFi association events. Each GPS location sample contains the timestamp, longitude, latitude, user ID, and error (in meters). Each AP association log contains the timestamp, the ID of the AP, and whether the event is association or dissociation. On Andriod phones, the GPS data were collected periodically every 10 minutes. On iPhones, there is no convenient mechanism for collecting GPS data periodically, and therefore we designed an event-based data method. For both platforms, the WiFi association data were logged based on events. See more details on data collection in [25].

GPS and WiFi data are complementary to each other: GPS works well outdoors while WiFi works better indoors; GPS provides finer-granularity location data than WiFi but is more energy consuming. We have developed a technique to fuse the two sources of data for more complete location coverage [87]. After fusion, the location data is represented as longitude and latitude pairs at the granularity of one minute; the time points with unknown locations are marked with unknown. The fused location data is used in the analysis.

3.3.2 Meta-data from WiFi Infrastructure

The meta-data from the WiFi infrastructure refers to the WiFi association logs that were captured at the APs (note that it differs from the WiFi association data in Section 4.3.1, which was collected on the phones, not from APs). Specifically, the information collected at the APs were queried by the university's IT services using standard network management protocols, and then sent to us on a regular basis. Each record corresponds to an AP association event, including the MAC address of the AP, the MAC address of the wireless device, the start time of the association, and the duration of the association. To preserve user privacy, for each AP association record, we hashed the MAC addresses of both the AP and device to 16 bytes each, and then stored the hashed values on our data collection server. Finding a participant's AP association records was based on the hashed MAC address of the participant's phone. The location information in an AP association record is represented by the ID of the AP, instead of longitude and latitude values as in location data collected on smartphone phones. We further leverage additional information from the IT services to map each AP to a building on campus. After that, the location information is represented as the building IDs.

Note that the data can only be collected when a participant is on campus and connected to the campus WiFi infrastructure. Since most students were not on campus during the holidays (Thanksgiving and Christmas) and breaks (spring, winter and summer breaks), our data analysis excluded those time periods.

3.3.3 Questionnaire Responses

In Phase I study, a participant filled in a PHQ-9 questionnaire during the initial assessment, and then every two weeks using an app that we developed. In Phase II study, following the suggestions from our study clinician, we switched from PHQ-9 to QIDS since it allows finer-grained labeling of depression symptoms and more frequent self-reports from participants. A participant filled in a QIDS questionnaire initially and every week using an app that we developed. Both the PHQ-9 and QIDS apps sent a notification to a participant to fill in the respective questionnaire on the due dates, and sent a reminder to the participant if a questionnaire was not filled in three days after the due date. The filled-in questionnaires were encrypted at the phone and then sent to our secure data collection server.

3.3.4 Clinical Assessment

Every participant was assessed by a clinician at the beginning of the study. Specifically, using an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and PHQ-9/QIDS evaluation, the clinician classified individuals as either depressed or non-depressed during the initial screening. A participant with a diagnosis of depression must participate in treatment to remain in the study. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician) to confirm their self-reported PHQ-9/QIDS scores with their verbal report during the meetings.

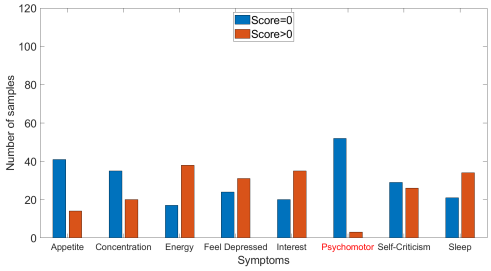
3.4 Predicting Depressive Symptoms Using Smartphone Sensing Data

In this section, we report the prediction results of individual depressive symptom using sensing data directly collected from smartphones; the results when using metadata collected from WiFi infrastructure are deferred to Section 3.5. In the following, we first describe data preprocessing and feature extraction. We then describe the classification methodology and the prediction results.

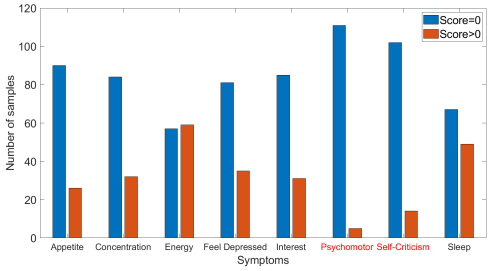
3.4.1 Data Preprocessing and Feature Extraction

We preprocess the data collected in PHQ-9 or QIDS intervals for each participant. Specifically, a *PHQ-9 interval* includes 15 days, the day when a PHQ-9 questionnaire is filled in and the previous 14 days (since PHQ-9 asks depressive symptoms in the previous two weeks). Similarly, a *QIDS interval* includes 8 days, the day when a QIDS is filled in and the previous 7 days (QIDS asks depressive symptoms in the previous week). Each PHQ-9/QIDS interval contains a questionnaire response from a participant, and the associated sensing data collected on smartphones. We only focus on location data, which is obtained by fusing the location data from GPS and WiFi association data that were collected on the phones (see Section 4.3.1). Even after data fusion, we still have substantial missing data. We therefore omit the PHQ-9/QIDS intervals with low data coverage in the data analysis. Specifically, if a PHQ-9 interval has less than 13 days with data or has less than 40% of the data points in the days with data, we omit the PHQ-9 interval in the analysis. For a QIDS interval, the corresponding thresholds are 6 days and 50%, respectively.

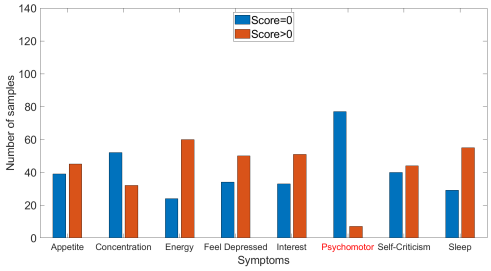
Samples. We now describe the number of samples (each corresponding to a self-



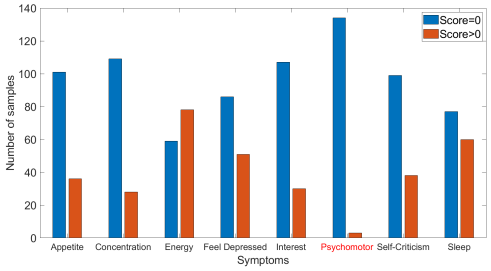
(A) Android, depressed.



(B) Android, non-depressed.

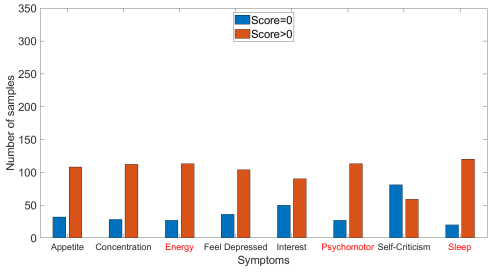


(C) iPhone, depressed.

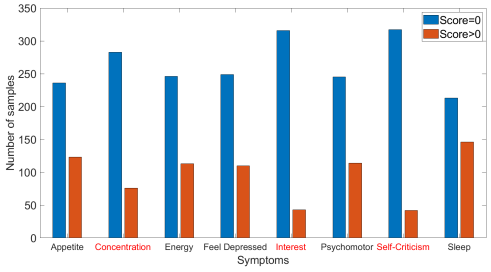


(D) iPhone, non-depressed.

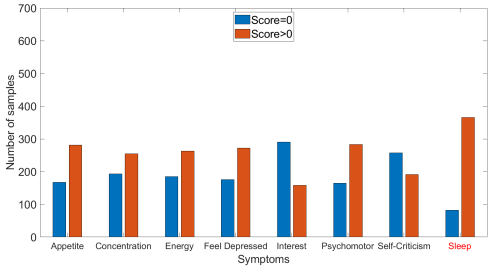
FIGURE 3.1: Number of samples for individual depressive symptoms (Phase I study).



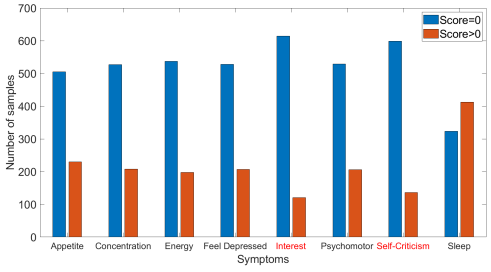
(A) Android, depressed.



(B) Android, non-depressed.



(C) iPhone, depressed.



(D) iPhone, non-depressed.

FIGURE 3.2: Number of samples for individual depressive symptoms (Phase II study).

report interval) after the above data preprocessing procedures. We present the samples for Android and iPhone users separately because the data collection follows significantly different methodologies (see Section 4.3.1), and hence our analysis is conducted for these two platforms separately. For Phase I, 25 are Android users (6 depressed and 19 non-depressed) and 54 are iPhone users (13 depressed and 41 non-depressed). For Phase II, 34 are Android users (12 depressed and 22 non-depressed) and 69 are iPhone users (27 depressed and 42 non-depressed).

Fig. 3.1 plots the histograms of the scores for each depressive symptom for Phase I participants, where the depression status (i.e., whether one is depressed or not) is based on clinician assessment. The score for a symptom is either 0 (no symptom) or larger than 0 (i.e., 1, 2, or 3, corresponding to slight, moderate or severe symptom, respectively). The reason for plotting only these two types of scores is that the number of samples of scores 2 and 3 is low, and hence we group them together with the samples with scores of 1. Our classification (Section 3.4.3) is therefore for two classes: absence and presence of symptom for each individual symptom. We see from the figure that for depressed participants, not surprisingly, a higher fraction of the scores for a depressive symptom is greater than 0 (i.e., has the symptom), while for non-depressed participants, a higher fraction of the scores is zero (i.e., does not have the symptom). A symptom label is marked in red if the samples are significantly unbalanced (specifically, the number of samples in one class is more than $4\times$ or less than 1/4 of the other), which will not be used in the analysis later on. The question of suicidal ideation is omitted from the figure since the scores are predominantly 0.

Fig. 3.2 plots the histograms of the scores for each depressive symptom for Phase II participants. We observe similar trends as those in Phase I. Again, the question of suicidal ideation is omitted due to the reason as described earlier.

Feature Extraction. We extract the following 10 features from the location data. The first four features are directly based on location data, while the last six features are based on locations clusters obtained using DBSCAN [21], a density based clustering algorithm to cluster the stationary points (i.e., those with moving speed less than 1km/h). DBSCAN requires two parameters, epsilon (the distance between points) and the minimum number of points that can form a cluster (i.e., the minimum cluster size). We varied the settings for these two parameters and selected the settings that led to the best overall correlations between the features and the PHQ-9/QIDS scores. For both Phases I and II, we set epsilon as 0.0002 (approximately 22 meters). For Phase I, the minimum number of points is set to correspond to 2.5 hours' stay (i.e., 160 since two adjacent locations are one minute apart after data fusion). For Phase II, it is set to correspond to around 3 hours' stay.

- **Location variance.** This feature [61] measures the variability in a participant's location. It is calculated as $\log(\sigma^2_{\text{long}}+\sigma^2_{\text{lat}})$, where σ^2_{long} and σ^2_{lat} represent the variance of the longitude and latitude of the location coordinates, respectively.

- **Time spent in moving.** This feature represents the percentage of time that a participant is moving. Specifically, as in [61], we estimate the moving speed at a sensed location, and treat a speed larger than 1km/h as moving, and as stationary otherwise.

- **Total distance.** Given the longitude and latitude of two consecutive location samples for a participant, we use Harversine formula [66] to calculate the distance traveled in kilometers between these two samples. The total distance traveled during a time period is the total distance normalized by the time period.

- **Average moving speed.** This feature represents the average moving speed, where movement and speed are identified in the same way as what is used for the total distance feature.

- **Number of unique locations.** It is the number of unique clusters from the DBSCAN algorithm.

- **Entropy.** It measures the variability of time that a participant spends at different locations. Let p_i denote the percentage of time that a participant spends in location cluster i. The entropy and is calculated as $-\sum(p_i \log p_i)$.

- **Normalized entropy.** It is entropy divided by the number of unique clusters. Hence it is invariant to the number of clusters and depends solely on the distribution of the visited location clusters [61].

- **Time spent at home.** This feature represents the percentage of time when a participant is at home. Following [61], we identify "home" for a participant as the location cluster that the participant is most frequently found between $[0, 6]$am.

- **Circadian Movement.** This feature is calculated as in [61]. It measures to what extent a participant's sequence of locations followed a 24-hour or circadian rhythm. To calculate circadian movement, we first use the least-squares spectral analysis, also known as the Lomb-Scargle method [55], to obtain the spectrum of the locations (represented by the cluster IDs). We then calculate the amount of energy that falls into the frequency bins within a 24 ± 0.5 hour period as

$$\mathrm{E} = \sum_i psd(f_i) / \ (i_1 - i_2)\,, \tag{3.1}$$

where $i = i_1, i_1 + 1, \ldots, i_2$, and i_1 and i_2 represent the frequency bins corresponding to 24.5 and 23.5 hour periods, respectively, $psd(f_i)$ denotes the power spectral density at each frequency bin f_i. The total circadian movement is then calculated as $\log(E)$.

- **Routine Index.** This feature is adapted from [13]. It quantifies how different the locations (represented by the cluster IDs) visited by a user in a day differs from those visited in another day. Specifically, it considers two days d_1 and d_2 in a self-report interval (i.e., PHQ-9 or QIDS interval). Let $\ell_{i1}, \ldots, \ell_{in}$ denote the locations that were visited in each minute on day i, $i = 1, 2$ (we only consider the set of intervals where there are recorded locations in both days). Then the similarity of these two days is

$$sim(d_1, d_2) = \left(\sum_{j=1}^{n} g(\ell_{1j}, \ell_{2j}) \right) / n \,, \tag{3.2}$$

where $g(\ell_{1j}, \ell_{2j}) = 1$ if $\ell_{1j} = \ell_{2j}$, and is zero otherwise. We see the value of $sim(d_1, d_2)$ is between 0 and 1, and a larger value represents a higher degree of similarity. The routine index of a self-report interval is the average of the similarities of all pairs of days within the interval. It is a value between 0 and 1; higher values indicate that the locations visited over the days are more similar.

3.4.2 Classification Methodology

For each depressive symptom, we used Support Vector Machine (SVM) models with a RBF kernel [14] for classifying whether one has the symptom or not. The classification is done for each self-report interval, using the self-report from a participant as the

ground truth. The presence of the depressive symptom is considered as positive and the absence of the symptom is considered as negative. The SVM model has two hyper-parameters, the cost parameter C and the parameter γ of the radial basis functions. We used leave-one-user-out cross validation procedure (i.e., no data from one user was used in both training and testing to avoid overfitting) to choose these two parameters. Specifically, we varied C and γ both in $2^{-15}, 2^{-14}, \ldots, 2^{14}, 2^{15}$, and chose the values that gave the best validation F_1 score. The F_1 score, defined as $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, is a weighted average of the precision and recall. It ranges from 0 to 1, and the higher, the better.

The above choice of parameters is performed for a given set of features. For each depressive symptom, we further selected the best set of features using SVM recursive feature elimination (SVM-RFE) [33, 57, 86], which is a wrapper-based feature selection algorithm designed for SVM. The goal of SVM-RFE is to find a subset of features out of all the features to maximize the performance of the SVM predictor. For a set of n features, we used SVM-RFE for feature selection as follows. For each pair of values for C and γ, SVM-RFE provided a ranking of the features, from the most important to the least important. After that, for each feature, we obtained its average ranking across all the combinations of C and γ values, leading to a complete order of the features. We then varied the number of features, k, from 1 to n. For a given k, the top k features were used to choose the parameters, C and γ, to maximize F_1 score based on the leave-one-user-out cross validation procedure as described above. The set of top k features that provides the highest F_1 score is chosen as the best set of features.

3.4.3 Symptom Prediction Results

We next present the prediction results of the individual depression symptoms for Phase I and Phase II studies. The results for the Android and iOS datasets are presented separately because of the significantly different data collection mechanisms that were used on these two platforms. For each platform, we report the results for three cases: all the participants, the depressed participants, and the non-depressed participants. The first case (all the participants) is interesting since it is for the (common) scenario when there is a mixture of depressed and non-depressed users, and the ground truth of the depression status is unknown. The second case is interesting since it provides insights on whether smartphone sensing data can be used by depressed users to categorize their individual depression symptoms automatically. Similarly, the third case provides insights on whether the automatic depression symptom categorization can be used by non-depressed users. The prediction results shown below include F_1 score as well as three other important performance metrics (including precision, recall, specificity). In addition, the number of features in the best set of features that was selected for the prediction is listed in the table. As mentioned earlier, the suicidal intent symptom is excluded from the analysis (since the responses are predominantly 0).

Phase I Results. Table 3.1 presents the classification results for Phase I study. For each depressive symptom, the number of samples from Android users is 171 (54 from depressed and 116 from non-depressed participants); the number of samples from the iOS users is 221 (84 from depressed and 137 from non-depressed participants). A symptom label is marked in red in Fig. 3.1 if the numbers of positive and negative samples are very unbalanced (specifically, their ratio is over 4 or less than 1/4). The

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Energy	0.69	0.66	0.73	0.50	10
		Feeling-depressed	0.70	0.66	0.74	0.53	4
		Interest	0.62	0.63	0.61	0.55	2
		Self-criticism	0.71	0.70	0.72	0.73	5
		Sleep	0.64	0.62	0.65	0.63	5
	Depressed	Energy	0.76	0.78	0.74	0.76	5
		Feeling-depressed	0.69	0.67	0.71	0.54	4
		Interest	0.69	0.63	0.77	0.60	6
		Sleep	0.65	0.57	0.76	0.52	5
	Non-depressed	Concentration	0.62	0.68	0.56	0.70	3
		Energy	0.71	0.65	0.78	0.56	9
		Feeling-depressed	0.70	0.65	0.76	0.66	3
		Interest	0.62	0.59	0.65	0.52	5
		Sleep	0.61	0.65	0.57	0.78	3
iOS	All	Appetite	0.75	0.74	0.75	0.70	7
		Concentration	0.69	0.69	0.69	0.65	3
		Energy	0.61	0.54	0.71	0.50	5
		Feeling-depressed	0.73	0.74	0.73	0.78	10
		Interest	0.76	0.80	0.73	0.79	7
		Self-criticism	0.80	0.81	0.79	0.78	6
		Sleep	0.71	0.71	0.70	0.69	8
	Depressed	Appetite	0.79	0.75	0.84	0.67	9
		Concentration	0.70	0.65	0.75	0.50	4
		Energy	0.74	0.64	0.87	0.57	5
		Feeling-depressed	0.83	0.90	0.76	0.88	5
		Interest	0.81	0.78	0.84	0.82	5
		Self-criticism	0.81	0.80	0.82	0.78	4
		Sleep	0.68	0.68	0.69	0.70	4
	Non-depressed	Energy	0.67	0.65	0.69	0.51	2
		Feeling-depressed	0.65	0.64	0.67	0.56	9

TABLE 3.1: Prediction of individual depressive symptoms for Phase I study (using smartphone sensing data).

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Concentration	0.60	0.66	0.54	0.66	5
		Interest	0.64	0.65	0.63	0.63	6
		Self-criticism	0.65	0.64	0.65	0.62	2
	Depressed	Appetite	0.68	0.59	0.81	0.53	10
		Concentration	0.63	0.62	0.63	0.61	3
		Feeling-depressed	0.65	0.61	0.68	0.58	2
		Interest	0.69	0.63	0.76	0.60	2
	Non-depressed	Feeling-depressed	0.60	0.63	0.54	0.72	3
		Psychomotor	0.60	0.56	0.60	0.56	2
iOS	All	Concentration	0.63	0.62	0.64	0.50	10
		Energy	0.63	0.63	0.63	0.52	2
		Feeling-depressed	0.67	0.66	0.68	0.52	4
		Self-criticism	0.61	0.59	0.63	0.50	10
		Sleep	0.64	0.63	0.66	0.62	10
	Depressed	Appetite	0.61	0.55	0.68	0.53	5
		Interest	0.60	0.53	0.69	0.50	4
	Non-depressed	Energy	0.60	0.58	0.59	0.52	8
		Feeling-depressed	0.64	0.61	0.66	0.50	7

TABLE 3.2: Prediction of individual depressive symptoms for Phase II study (using smartphone sensing data).

results below exclude such symptoms. For the remaining symptoms, we only list the symptoms with the resultant F_1 score above 0.6 in the table.

We first describe the results for Android users. As shown in Figures 3.1 (a) and (b), psychomotor is excluded from the analysis for both depressed and non-depressed participants, and in addition self-criticism is excluded from the analysis for non-depressed participants. From Table 3.1 (the top part), we see that for depressed participants, four symptoms, energy level, feeling depressed, interest and sleep disturbance, were predicted with significant F_1 scores (above 0.6). These four symptoms were also predicted with significant F_1 scores for all and non-depressed participants. In addition, another symptom, self-criticism, was predicted accurately for all participants; and concentration was predicted accurately for non-depressed participants.

The above results show that, maybe surprisingly, location features, which are behavioral in nature, can be used to predict cognitive symptoms such as feeling de-

pressed, interest and concentration accurately. This may be because location characteristics are inherently correlated with cognitive symptoms. For instance, feeling depressed or lack of interests may lead one to move less, visit less places and stay in a smaller number of places for a longer period of time. Indeed, the features that were selected for predicting feeling depressed include entropy, normalized entropy, number of locations, and distance traveled. The features that were selected for predicting interests and concentration include location variance, entropy, normalized entropy, the amount of time moving, and the distance traveled. We further observe that sleep can be predicted accurately, consistent with existing studies that show that smartphones data can be used for detecting sleeping patterns [51, 16, 52, 34]. We also observe that, for the symptoms that were predicted accurately for depressed participants, their F_1 scores tend to be higher than the corresponding F_1 scores for the other two cases (i.e., all the participants and the non-depressed participants), indicating that the smartphone sensing data is more effective in keeping track of the depression symptoms for depressed participants. This might be because, for depressed participants, their self-report scores of the individual symptoms reflect more consistently their psychological status. This observation is consistent with results in our prior work [49, 83], which show that location features are more correlated with the overall self-report scores (i.e., the sum of the scores of the individual symptoms) for depressed participants than that for the non-depressed participants.

Table 3.1 (bottom part) shows the results for iOS users. As shown in Figures 3.1 (c) and (d), psychomotor is excluded from the analysis due to significantly unbalanced samples. We see that all seven symptoms (i.e., all the nine symptoms excluding psychomotor and suicidal intent) that we considered were predicted accurately for depressed users. The F_1 score ranges from 0.61 to 0.83. Out of the seven symptoms, four

symptoms (concentration, feeling-depressed, interests and self-criticisms) are cognitive, confirming our earlier observation that location features can be used to predict cognitive symptoms accurately. For non-depressed participants, two symptoms, energy level and feeling depressed, were predicted accurately. For all the participants, all seven symptoms were predicted accurately, with the predicted F_1 scores slightly lower than those for the depressed participants, consistent with the results for the Android users.

Phase II Results. Table 3.2 presents the classification results for Phase II dataset. For each depressive symptom, we have 499 samples from the Android users (140 from depressed and 359 from non-depressed participants), and 1183 samples from the iOS users (448 from depressed and 735 from non-depressed participants). The number of samples in each case is significantly larger than that in Phase I study because each sample corresponds to a one-week time period (since the QIDS questionnaire used in Phase II asks about the symptoms in the past week), instead of two-week time period as in Phase I study. In addition, the number of participants (particularly, depressed participants) in Phase II is larger than that in Phase I.

The top part of Table 3.2 shows the results for Android participants. For the depressed participants, in addition to suicide intent, three symptoms (energy level, psychomotor, and sleep) were not considered in the analysis due to significantly unbalanced samples. Of the five remaining symptoms, four symptoms, one behavioral (appetite) and three cognitive (concentration, feeling depressed, and interest) symptoms, were predicted accurately; and only one symptom (self-criticism) was not predicted accurately. For the non-depressed participants, in addition to suicide intent, three symptoms (concentration, interests, and self-criticism) were excluded from the

analysis. Of the five remaining symptoms, two symptoms (feeling depressed and psychomotor) were predicted accurately. For all the participants, the symptoms that were predicted accurately include concentration, interest and self-criticism. We again observe higher F_1 scores for the depressed participants than those for the non-depressed and all participants.

The bottom part of Table 3.2 shows the results for iOS participants. For the depressed participants, two symptoms (suicide intent and sleep) were excluded from the analysis. Of the remaining seven symptoms, two symptoms (appetite and interest) were predicted with F_1 scores larger than 0.6. For non-depressed participants, two symptoms (energy and feeling depressed) were predicted with significant F_1 scores. The fewer symptoms that were predicted accurately compared to Phase I results (see the bottom part of Table 3.1) might be due to two factors: (1) the location features in Phase II were extracted from one-week location data (instead of two-week data as in Phase I), and (2) the location data were collected using an event-based mechanism on iOS platform. While the first factor also holds true for Android data, the periodic data collection on Android leads to better location coverage than the event-based location collection on iOS platform [89], leading to impact on the classification results for individual symptoms. We believe that the Phase II iOS results can be improved by better data preprocessing (e.g., using better techniques for handling the missing data to provide more complete data coverage) and feature extraction (e.g., including more features); further investigation is left as future work. For all the participants, five symptoms were predicted accurately, which were predicted accurately in Phase I study as well.

Summary. Summarizing the results in Phase I and Phase II studies, we observe that

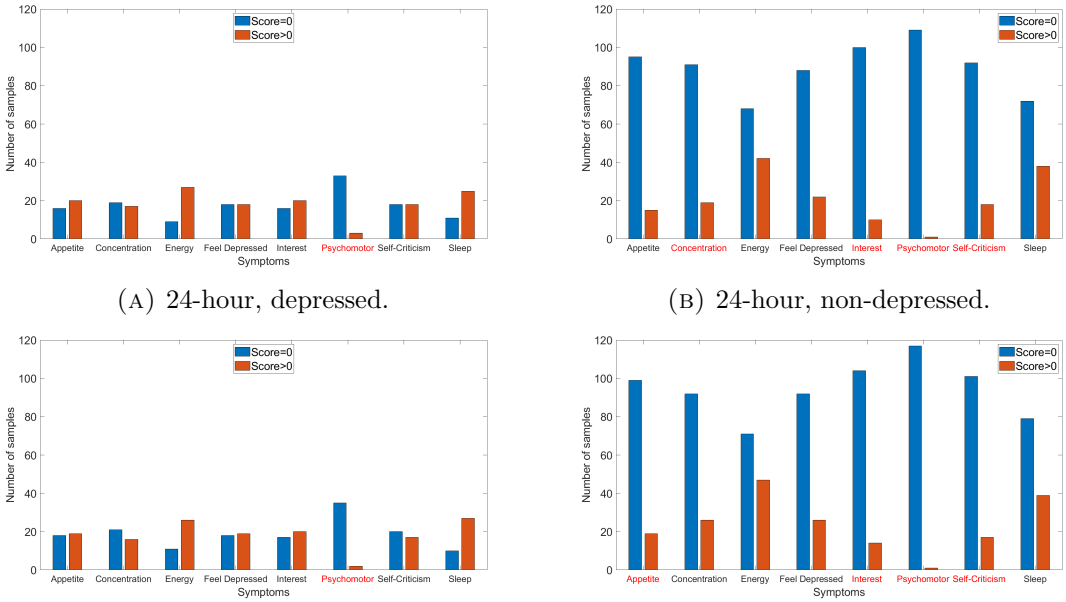
three symptoms, appetite, interest and feeling depressed, were predicted accurately across multiple settings for depressed participants. For non-depressed participants, one symptom, feeling depressed, was predicted accurately in all the settings. For all the participants, concentration, energy, feeling-depressed, interest, self-criticism, and sleep were predicted accurately in various settings. The predicted F_1 score is up to 0.83, comparable to the F_1 scores obtained for predicting the overall depression status [25, 87, 49, 89]. Given that the data was collected passively without any efforts from the users, automatic prediction using the collected data provides an attractive approach for keeping track of the absence and presence of depressive symptoms continuously over time. The differences in Phase I and II results, particularly for iOS users, also point out future directions in improving data collection, preprocessing and feature extraction to tackle the challenges of missing data.

3.5 Predicting Depressive Symptoms Using WiFi Infrastructure Data

We now present the prediction results of individual depressive symptom when using meta-data collected from WiFi infrastructure. In the following, we first describe data preprocessing and feature extraction, and then the prediction results. The classification methodology is the same as that presented in Section 3.4.2.

3.5.1 Data Preprocessing and Feature Extraction

As mentioned in Section 3.3.2, the location information is represented as building IDs (by mapping an AP that a phone is associated with to the building that the AP



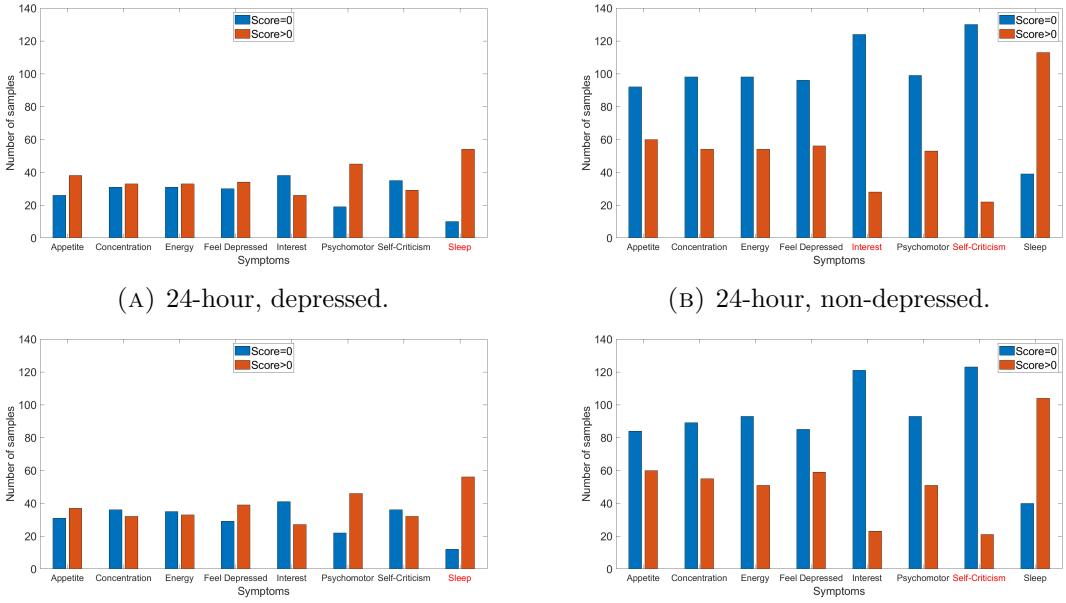
(A) 24-hour, depressed.

(B) 24-hour, non-depressed.

(C) daytime, depressed.

(D) daytime, non-depressed.

FIGURE 3.3: Number of samples for depressive symptoms for Phase I study (using WiFi infrastructure data).



(A) 24-hour, depressed.

(B) 24-hour, non-depressed.

(C) daytime, depressed.

(D) daytime, non-depressed.

FIGURE 3.4: Number of samples for depressive symptoms for Phase II study (using WiFi infrastructure data).

is located in). We again filter out PHQ-9 and QIDS-intervals that do not contain sufficient amount of data. Specifically, we exclude a PHQ-9 interval if it contains less than 14 days of data, and exclude a QIDS interval if it contains less than 7 days of data. Figures 3.3 and 3.4 plot the number of samples (PHQ-9/QIDS intervals) after the above data preprocessing for Phase I and Phase II studies, respectively. The results for the two scenarios, 24-hour and daytime monitoring, are shown in the figures. We again plot the scores for the depressed and non-depressed participants separately (the depression status is based on clinician assessment). As expected, for each symptom, the fraction of the scores larger than 0 (i.e., with the symptom) is higher for the depressed participants than that for the non-depressed participants.

We extracted the following 18 features from the data. The first six features, including the number of unique locations, entropy, normalized entropy, time spent at home (only applicable to 24-hour monitoring), circadian movement, routine index, are the same as those used for smartphone sensing data (Section 3.4.1), except that the locations are represented as building IDs instead of cluster IDs. The remaining 12 features are described below; most of them are related to the categories of the buildings (we broadly classified the campus buildings based on their main purposes as entertainment, sports, class, library, and others).

- **Number of significant locations visited**. This featured is calculated as in [13]. Let S denote the top 10 most significant buildings visited by a user (i.e., the 10 buildings where a user spent the most time) during the period of study. The number of significant locations in a self-report interval (i.e., PHQ-9 or QIDS interval) is the number of unique buildings visited in the interval that are in S.

- **Number of Entertainment, Sports and Class buildings visited.** The campus has multiple entertainment, sports and class buildings. For each category of buildings, we calculated the number of unique buildings visited by a participant in a given PHQ-9 or QIDS interval.

- **Average duration spent in Entertainment, Sports, Library and Class buildings.** These features represent the average duration that a participant spent in each category of buildings over a PHQ-9 or QIDS interval.

- **Number of days visiting Entertainment, Sports, Library and Class buildings.** These features represent the number of days that a participant visited a specific category of buildings over a PHQ-9 or QIDS interval.

3.5.2 Symptom Prediction Results

We next present the prediction results. For both 24-hour and daytime monitoring, we again report the results for three cases: all the participants, the depressed participants, and the non-depressed participants. Again, the suicidal intent symptom is excluded from the analysis.

Phase I Results. Table 3.3 presents the classification results for Phase I study. For each symptom, the number of samples from all the participants is 146 for 24-hour monitoring (36 from depressed and 110 from non-depressed participants); the number of samples from all the participants is 155 for daytime monitoring (37 from depressed and 118 from non-depressed participants). Again, a symptom marked in red in Fig. 3.3 is excluded from the analysis due to significantly unbalanced samples.

We first present the results for depressed participants. For both 24-hour and day-

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-hour monitoring	All	Appetite	0.79	0.78	0.80	0.79	7
		Concentration	0.60	0.60	0.58	0.62	4
		Energy	0.63	0.61	0.67	0.61	13
		Feeling-depressed	0.71	0.73	0.70	0.71	18
		Interest	0.72	0.65	0.80	0.66	2
		Self-criticism	0.64	0.62	0.67	0.59	6
		Sleep	0.67	0.65	0.68	0.72	9
	Depressed	Appetite	0.76	0.68	0.85	0.50	2
		Concentration	0.63	0.57	0.71	0.53	5
		Energy	0.80	0.73	0.89	0.67	3
		Feeling-depressed	0.85	0.83	0.78	0.84	2
		Interest	0.86	0.79	0.85	0.56	3
		Self-criticism	0.86	0.84	0.89	0.83	2
		Sleep	0.70	0.65	0.76	0.54	3
	Non-depressed	Sleep	0.73	0.71	0.76	0.67	4
Daytime monitoring	All	Appetite	0.63	0.66	0.61	0.70	11
		Concentration	0.68	0.63	0.74	0.51	5
		Energy	0.68	0.68	0.68	0.71	7
		Feeling-depressed	0.61	0.74	0.51	0.85	2
		Interest	0.69	0.67	0.71	0.61	5
		Self-criticism	0.66	0.75	0.59	0.78	9
		Sleep	0.60	0.62	0.58	0.74	17
	Depressed	Appetite	0.84	0.84	0.83	0.83	2
		Concentration	0.72	0.61	0.88	0.57	3
		Energy	0.68	0.68	0.68	0.70	7
		Feeling-depressed	0.82	0.72	0.90	0.61	3
		Interest	0.76	0.73	0.80	0.65	7
		Self-criticism	0.85	0.88	0.82	0.90	8
		Sleep	0.75	0.69	0.81	0.50	2
	Non-depressed	Sleep	0.68	0.68	0.69	0.67	8

TABLE 3.3: Prediction of individual depressive symptoms for Phase I (using WiFi infrastructure data).

time monitoring, in addition to suicidal intent, psychomotor was excluded from the analysis. All the seven remaining features were predicted with significant F_1 scores. Specifically, all four cognitive symptoms (concentration, interest, feeling depressed, self-criticism) were predicted accurately using location features, consistent with earlier prediction results using smartphone sensing data (Section 3.4.3). The number of selected features tends to be small. The selected features include (normalized) entropy, circadian movement, routine index and various features related to building semantics (e.g., number of days of visiting sports buildings, number of days or durations visiting library buildings).

For all the participants, we again observe that all the seven symptoms that were analyzed were predicted accurately. The F_1 scores are slightly lower than those for depressed participants, consistent with the prediction results when using smartphone sensing data (Section 3.4.3). The number of selected features is large for some symptoms. For non-depressed participants, four symptoms were considered in analysis (the other five symptoms do not have sufficiently balanced samples) for both 24-hour and day-time monitoring. Out of these four symptoms, sleep was predicted accurately.

Phase II Results. Table 3.4 presents the classification results for Phase II study. For each symptom, the number of samples from all the participants is 216 for 24-hour monitoring (64 from depressed and 152 from non-depressed participants); the number of samples from all the participants is 212 for daytime monitoring (68 from depressed and 144 from non-depressed participants).

We again first present the results for depressed participants. For both 24-hour and daytime monitoring, sleep and suicidal intent were excluded from the analysis. All the seven remaining symptoms were predicted accurately using 24-hour monitoring;

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-hour monitoring	All	Appetite	0.68	0.60	0.79	0.56	2
		Concentration	0.63	0.65	0.61	0.78	2
		Energy	0.65	0.71	0.60	0.84	7
		Interest	0.65	0.68	0.63	0.70	6
		Psychomotor	0.64	0.61	0.68	0.64	16
		Sleep	0.72	0.65	0.81	0.51	7
	Depressed	Appetite	0.85	0.79	0.90	0.65	7
		Concentration	0.75	0.82	0.70	0.84	4
		Energy	0.77	0.82	0.73	0.84	4
		Feeling-depressed	0.67	0.63	0.71	0.53	2
		Interest	0.65	0.65	0.65	0.76	11
		Psychomotor	0.76	0.68	0.87	0.53	7
		Self-criticism	0.62	0.70	0.55	0.80	3
	Non-depressed	Concentration	0.78	0.80	0.76	0.80	7
		Energy	0.79	0.80	0.78	0.78	7
		Feeling-depressed	0.69	0.69	0.69	0.63	2
		Psychomotor	0.69	0.64	0.75	0.55	2
		Sleep	0.66	0.65	0.66	0.74	5
Daytime monitoring	All	Appetite	0.64	0.65	0.64	0.70	12
		Concentration	0.62	0.71	0.55	0.84	6
		Psychomotor	0.63	0.58	0.68	0.58	3
		Self-criticism	0.65	0.61	0.70	0.55	4
		Sleep	0.71	0.67	0.75	0.62	7
	Depressed	Appetite	0.69	0.68	0.70	0.61	6
		Concentration	0.67	0.65	0.69	0.67	3
		Energy	0.70	0.67	0.73	0.66	6
		Feeling-depressed	0.71	0.67	0.74	0.51	7
		Interest	0.67	0.72	0.63	0.68	5
		Psychomotor	0.62	0.58	0.67	0.50	3
	Non-depressed	Concentration	0.77	0.73	0.82	0.62	5
		Energy	0.72	0.71	0.73	0.68	5
		Psychomotor	0.67	0.65	0.69	0.60	6
		Sleep	0.70	0.62	0.80	0.58	4

TABLE 3.4: Prediction of individual depressive symptoms for Phase II (using WiFi infrastructure data).

for daytime monitoring, all seven symptoms except for self-criticism were predicted accurately. For all participants, six and five symptoms were predicted accurately for 24-hour and daytime monitoring, respectively. For non-depressed participants, of the six symptoms considered in the analysis for 24-hour monitoring, five were predicted accurately; for daytime monitoring, four out of the seven symptoms were predicted accurately. Again, the predicted F_1 scores for depressed participants tend to be higher than those for all the participants and non-depressed participants.

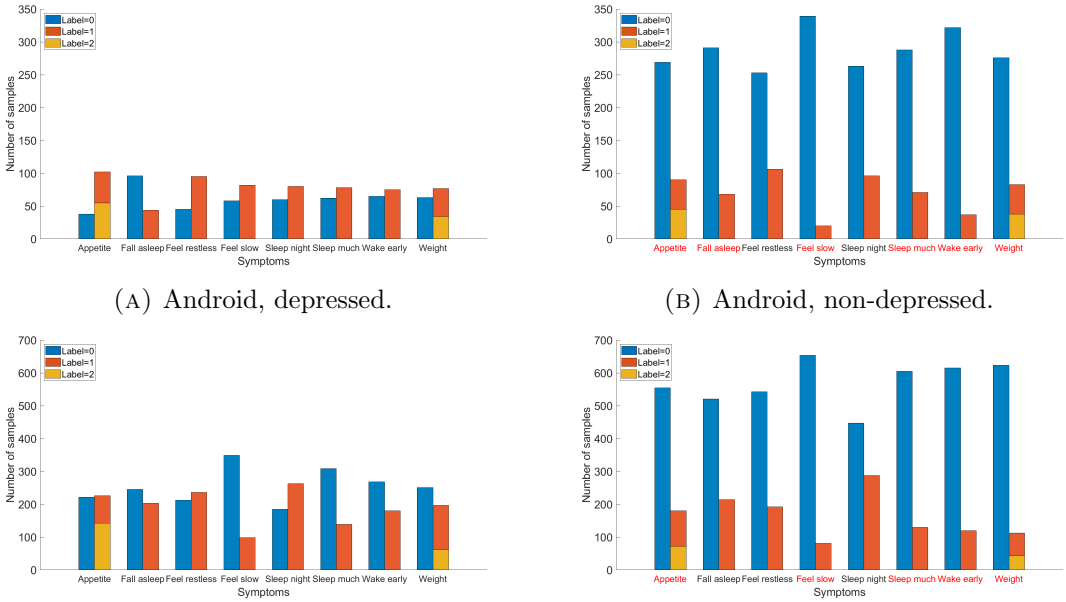
		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
Android	All	Feeling slowed down	0.67	0.71	0.64	0.73	5
		Sleeping too much	0.60	0.59	0.58	0.65	3
	Depressed	Falling asleep	0.68	0.61	0.77	0.56	2
		Feeling restless	0.68	0.63	0.73	0.56	3
		Sleep during the night	0.72	0.69	0.76	0.56	2
		Sleeping too much	0.66	0.66	0.66	0.57	3
		Waking up too early	0.68	0.65	0.71	0.55	3
	Non-Depressed	Feeling restless	0.60	0.58	0.61	0.62	2
iOS	All	Falling asleep	0.66	0.65	0.67	0.62	5
		Feeling restless	0.64	0.61	0.67	0.52	10
		Waking up too early	0.63	0.60	0.67	0.53	2
	Depressed	Sleep during the night	0.68	0.67	0.69	0.51	2
	Non-Depressed	Sleep during the night	0.62	0.62	0.63	0.50	5

TABLE 3.5: Prediction of finer-level individual depressive symptoms (Phase II, using smartphone sensing data).

Summary. The above results of both Phase I and Phase II studies demonstrate that location features automatically extracted from WiFi network meta-data can be used to predict individual depressive symptoms accurately, even in the cases where only daytime information is available. The prediction is more effective for depressed participants. For the overall population, a wide range of depressive symptoms can be predicted with good accuracy. The predicted F_1 score is up to 0.86, comparable to that for predicting the overall depression status [83]. The prediction will be helpful for an institution to keep tabs on the overall mental health status of the employees/students in the institution at very little cost. We further found that the features related to the building categories are particularly useful for classification, highlighting the importance of including fine-grained location features.

3.6 Predicting Finer-level Depressive Symptoms

In QIDS questionnaire (used in Phase II study), some symptoms have multiple sub-questions, presenting finer-level symptom information. Specifically, there are four sub-



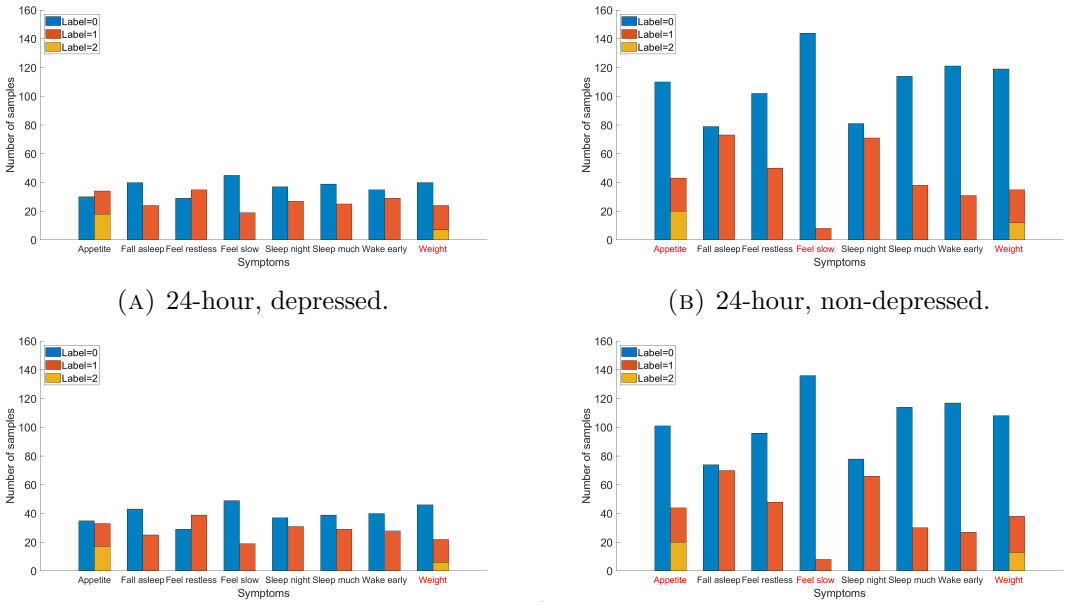
(A) Android, depressed.

(B) Android, non-depressed.

(C) iPhone, depressed.

(D) iPhone, non-depressed.

FIGURE 3.5: Number of samples for finer-grain depressive symptoms for Phase II study (corresponding to smartphone sensing data).



(A) 24-hour, depressed.

(B) 24-hour, non-depressed.

(C) daytime, depressed.

(D) daytime, non-depressed.

FIGURE 3.6: Number of samples for finer-grain depressive symptoms for Phase II study (using WiFi infrastructure data).

questions on sleep disturbance, four sub-questions on appetite/weight, and two sub-questions on psychomotor agitation/retardation. We next explore using smartphone sensing data and WiFi infrastructure data to predict these finer-grain depressive symptoms (see Section 3.2.1).

Methodology. Figures 3.5 and 3.6 plot the number of samples (QIDS intervals) for each finer-level depressive symptom. For most symptoms, the label is either 0 (no symptom, i.e., score is 0) or 1 (with symptom, i.e., the score is larger than 0). For appetite and weight, the label is 0 (no change in appetite/weight), 1 (increased appetite/weight), or 2 (decreased appetite/weight) since the sub-questions on increased or decreased appetite/weight are mutual exclusive. Again we see that non-depressed participants have significantly higher fraction of samples with label 0 (i.e., no symptom) than depressed participants. We used the same features as described in Sections 3.4 and 3.5 for prediction. The classification methodology for two-label symptoms is as described in Section 3.4.2. For the symptoms with three labels, we again used SVM models; the hyper-parameters were chosen so that the average F_1 scores of the three classes is maximized.

Prediction Results. Table 3.5 presents the classification results using smartphone sensing data. The results for both Android and iOS users were listed in the table. As observed in Section 3.4.3, the prediction results for Android users were (slightly) better than those of the iOS users, which may be due to more complete data on Android phones. The F_1 scores for appetite and weight were below 0.6 and not presented in the table. The F_1 scores for some sleep related finer-level symptoms were significant. Given the intricate relationship between sleep and depression [74, 9, 84], treating sleep disorders is an important component of treating depression.

		Depressive symptom	F_1 Score	Precision	Recall	Specificity	# of features selected
24-hour monitoring	All	Falling asleep	0.61	0.55	0.66	0.56	2
		Feeling restless	0.75	0.70	0.80	0.56	13
		Sleep during the night	0.60	0.57	0.63	0.61	7
		Sleeping too much	0.70	0.70	0.70	0.76	8
		Waking up too early	0.62	0.60	0.65	0.50	2
	Depressed	Falling asleep	0.83	0.79	0.88	0.73	10
		Feeling slowed down	0.64	0.71	0.58	0.80	6
		Feeling restless	0.62	0.61	0.63	0.52	5
		Sleep during the night	0.60	0.70	0.52	0.84	4
		Sleeping too much	0.86	0.85	0.88	0.79	4
		Waking up too early	0.72	0.72	0.72	0.77	3
	Non-Depressed	Falling asleep	0.67	0.65	0.70	0.65	9
		Feeling restless	0.65	0.61	0.70	0.56	10
		Sleep during the night	0.68	0.63	0.75	0.62	2
		Sleeping too much	0.77	0.84	0.71	0.86	8
Daytime monitoring	All	Falling asleep	0.61	0.57	0.65	0.60	3
		Sleep during the night	0.65	0.58	0.73	0.55	4
		Sleeping too much	0.73	0.66	0.81	0.52	2
		Waking up too early	0.68	0.64	0.62	0.64	2
	Depressed	Feeling restless	0.65	0.66	0.64	0.55	4
		Feeling slowed down	0.71	0.74	0.68	0.82	3
		Sleep during the night	0.67	0.66	0.68	0.70	8
		Sleeping too much	0.79	0.81	0.76	0.87	5
		Waking up too early	0.75	0.71	0.79	0.78	4
	Non-Depressed	Falling asleep	0.70	0.62	0.80	0.54	2
		Feeling restless	0.65	0.61	0.69	0.56	4
		Sleep during the night	0.66	0.61	0.71	0.62	3
		Sleeping too much	0.70	0.68	0.73	0.63	2

TABLE 3.6: Prediction of finer-level individual depressive symptoms (Phase II, using WiFi infrastructure data).

Our results indicate that smartphone data can be used to automatically keeping track of sleeping disorders, particularly for depressed patients during treatment, and potentially providing objective assessment on whether the treatment is effective or not for a patient.

Table 3.6 lists the classification results using meta-data collected from WiFi infrastructure. Again, the F_1 scores for appetite and weight were insignificant (and hence not included in the table). On the other hand, of the four sleep related finer-level symptoms and the two psychomotor related finer-level symptoms, most were predicted with significant F_1 scores, with the F_1 scores for depressed participant higher than those for the other two cases (i.e., all and non-depressed participants). Comparing the results in Tables 3.5 and Table 3.6, we see that the prediction results when using WiFi infrastructure data are better than those using the smartphone sensing data. The latter can be improved by further refining feature extraction and data collection, which is left as future work.

3.7 Related Work

Recent studies have used smartphone sensing data for depression prediction [28, 31, 32, 13, 81, 61, 6, 50, 91, 80, 54, 24, 25, 71, 19, 87, 49]. Wang et al. [81] reported a significant correlation between depressive mood and social interaction (specifically, conversation duration and number of co-locations). Saeb et al. [61] found significant correlation between sensing features (phone usage and mobility patterns) and the self-reported PHQ-9 scores. Canzian and Musolesi [13] studied the relationship between the mobility patterns and depression, and found that individualized machine learning

models outperformed general models. Farhan et al. [25] found that the features extracted from the smartphone sensing data can predict depression with good accuracy. Suhara et al. [71] developed a deep learning based approach that forecasts severely depressive mood based on self-reported histories. Yue et al. [87, 89] investigated fusing GPS and WiFi association data, both collected locally on smartphones, for more complete location information for improved depression detection. Lu et al. [49] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. Ware et al. [83] investigated the feasibility of using meta-data from WiFi infrastructure for automatic depression screening. Our study differs from the above studies in that we use smartphone data to predict individual depressive symptoms, instead of depression status (i.e., whether one is depressed or not). In addition to the nine broad categories of depressive symptoms, we further develop prediction models for finer-grain depressive symptoms.

As other related work, Torous et al. [73] investigated adherence among psychiatric outpatients diagnosed with major depressive disorder in utilizing their personal smartphones to run a custom app to monitor PHQ-9 depression symptoms. In addition, the authors examined the correlation of these scores with traditionally administered (paper-and-pencil) PHQ-9 scores. Simon et al. [69] found that the response to suicidal intent question in the questionnaire identifies patients at high risk of suicide attempt and suicide death. The studies in [51, 16, 52, 34] investigated using smartphone data to predict sleep qualities.

3.8 Conclusion and Future Work

In this part of the dissertation, we have investigated the feasibility of using smartphone data to predict individual depressive symptoms. We have constructed a family of machine learning based models that use features extracted from two types of smartphone data (i.e., smartphone sensing data and WiFi infrastructure meta-data) for the prediction. Our results, using data collected from 182 college students, demonstrated that a rich set of depressive symptoms can be predicted accurately using smartphone data. Furthermore, even finer-level depressive symptoms can be predicted accurately. Our study makes an important step forward over existing studies in demonstrating that using passively collected smartphone data is a promising direction in automatically keeping track of depressive symptoms. We primarily used location data in this work. The participants of the study were college students. Future directions include (1) using other types of sensing data (e.g., activity, SMS and email logs, web browsing records), and (2) exploring in other demographic groups.

Acknowledgments

We would like to thank all the participants who participated in our studies. We would also like to thank University of Connecticut Information Technology Services for providing us the WiFi infrastructure meta-data, and Prof. Shengli Zhou (UConn) for helpful discussions. This work was supported by the National Science Foundation (NSF) grant IIS-1407205. Jinbo Bi was also supported by NSF grants DBI1356655, CCF-1514357 and IIS-1718738, and NIH grants 5R01DA037349-04 and 5K02DA043063-03.

Chapter 4

Automatic Depression Screening Using Social Interaction Data

4.1 Introduction

Depression is a prevalent mental health problem that impacts the overall health of an individual, and incurs higher medical costs and mortality [20, 40, 68]. According to a recent national survey, an estimated 17.3 million adults (aged 18 or older) in the United States had at least one major depressive episode; this number represented 7.1% of all U.S. adults, and a major depressive episode was highest among individuals aged 18-25 (13.1%) [2].

Current diagnosis methods of depression are either clinician administered or patient self-administered. Such methods are often burdensome and not suitable for continuous monitoring. With the emergence of mobile computing and ubiquitous adoption of smartphones, recent studies have proposed novel approaches that use

smartphone sensing data for automatic depression screening (see Section 4.2). The intuition is that smartphones are equipped with a rich set of sensors (e.g., GPS, WiFi, activity, light); sensing data captured by these sensors can be used to derive meaningful features that indicate behavioral patterns of a person, e.g., the number of places visited, activity levels, etc. Such behavioral features can then be fed into machine learning algorithms (with pre-trained machine learning models) to automatically detect depression.

Most existing studies, however, focus on using physical location and activity data for depression screening, with little or no attention to social interaction data. Social interaction plays a significant role in the day-to-day life of an individual, with the social groups including family, friends, work place colleagues and others. Social ties are beneficial to psychological well-being [41], and both quality and quantity of social relationships affect mental health and mortality risk [76]. In this work, we investigate using social interaction data passively collected from smartphones for automatic depression prediction. Specifically, we focus on SMS messages and phone calls since they are two dominant modes of social communication, and play a central role in maintaining one's social networks [46, 35]. To preserve user privacy, we only consider statistical information related to the timing, frequency, quantity and variability of SMS and phone call activities; the content was never recorded. Both SMS and phone call logs can be easily captured with little energy consumption, much lower than that needed for capturing location and activity data.

While SMS and phone call logs have been used in several existing studies to correlate with or infer mental health states such as mood [47, 65], stress [63, 7], happiness [8], bipolar symptoms [27], or depression [58], these studies used SMS and phone call data together with a large variety of other types of sensing data (e.g., location,

activity, proximity). In contrast, we investigate using SMS or phone call logs alone, without any other type of sensing data, for depression prediction. Our focus on using a single type of data is advantageous in scenarios where only one type of data is available. For instance, in certain scenarios, due to energy consumption considerations, privacy issues or missing data, only SMS or phone call logs are collected successfully. In addition, our study quantifies the effectiveness of using SMS or phone call logs alone as a proxy of social interaction for depression prediction; including other types of data (when available) can further improve prediction accuracy.

Specifically, in this study, we collected SMS and phone call logs from 59 college students, all using Android phones; the depression status of each participant is based on clinician assessment. Using the data, we compare the characteristics of SMS and phone call logs for depressed and non-depressed participants. Furthermore, we investigate multiple classification methods that use SMS and phone call logs separately for depression prediction. Our study makes the following contributions:

- We extract a wide array of features from SMS and phone call logs, including characteristics on quantity, timing, length/duration, and variability (including standard deviation and entropy values) of these activities. Statistical analysis indicates that depressed and non-depressed participants have different characteristics in SMS and phone call usage patterns. Specifically, their behaviors in outgoing messages and phone calls (i.e., initiated by the users) are more statistically distinct than incoming messages and phone calls.

- Using the extracted features, we investigate using multiple machine learning models, including Support Vector Machine (SVM) [14], random forest [10] and XGBoost [15], for depression prediction. Our results show that XGBoost clas-

sifier achieves the best prediction. When using SMS logs, the highest predicted F_1 score is 0.80; when using phone call logs, the highest predicted F_1 score is 0.78. Both results are comparable to those achieved by using location and activity data [61, 13, 25, 89], demonstrating that SMS and phone call logs alone can already provide accurate depression prediction.

The rest of the chapter is organized as follows. Section 4.2 describes related work. Section 4.3 describes data collection. Section 4.4 describes feature extraction. Section 4.5 compares the characteristics of SMS and phone call logs of depressed and non-depressed participants. Section 4.6 presents depression prediction results when using SMS and phone call logs, respectively. Finally, Section 4.7 concludes the chapter, and briefly describes limitation of this study and future work.

4.2 Related Work

There are a large number of studies that use smartphone sensing data for mental health applications [28, 31, 32, 81, 61, 6, 50, 91, 80, 54, 24, 19]. In the following, we brief review related work in two directions, one using SMS and phone call logs, and the other using other types of sensing data.

Using SMS and phone call logs for mental health applications. Several studies used SMS and phone call logs for mental health applications. LiKamWa et al. [47] used a wide variety of mobile phone sensing data, including email, SMS, phone call logs, website domains, location clusters, apps, and categories of apps, to infer daily mood. They were able to infer a user's daily mood with an initial accuracy of 66% followed by improved accuracy of 93% after two months personalized training.

Mood prediction was also studied in [65], which was in a much larger scale (involving ∼18,000 users), again using a variety of smartphone data, ranging from physical activity, sociability, to mobility data, where sociability data included information on SMS and phone logs. Their results showed that especially on weekends, mobile sensing can be used to predict users' mood with an accuracy of about 70%. Bogomolov et al. [7] showed that daily stress can be reliably recognized based on behavioral metrics, derived from mobile phone usage patterns (including those extracted from SMS and phone call logs, and Bluetooth proximity data), and additional indicators, such as the weather conditions and personality traits. Their multifactorial statistical model obtained accuracy up to 72.28% for a 2-class daily stress recognition problem. The same types of data were used in [8] to recognize daily happiness. Sano and Picard [63] aimed to find physiological or behavioral markers for stress. They collected a large amount of data from wearable wrist sensors (accelerometer and skin conductance) and mobile phones (phone call and SMS logs, location and screen on/off) and various self-report surveys (stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake and electronics usage). Their results showed above 75% accuracy in a binary classification on stress using the various data. Faurholt-Jepsen et al. [27] studied the correlation between various phone sensing data (including phone usage such as screen on/off, changes in cellular tower, and social activities such as the number of incoming/outgoing messages and phone calls) with symptoms during depressive and manic periods for bipolar patients. Razavi et al. [58] examined the possibility of depression screening using mobile phone usage patterns, including daily mobile usage, basic characteristics of phone calls and text messages, amount of time spent on web browsing, social media and entertainment apps, and the number of saved contacts on device. The best model was a random forest classifier that had

an out-of-sample balanced accuracy of 76.8%, which was improved to 81.1% when including participants' age and gender information.

All the above studies used SMS and phone call logs together with a range of other sensing data (e.g., location, email, apps, websites, screen on/off), while our study only uses SMS or phone call logs alone. In the above, only the study in [58] considered depression, and hence is closest to our study. On the other hand, it differs from our study in several important aspects. In [58], participants were recruited through Amazon Mechanical Turk. The authors used Beck Depression Inventory 2nd edition (BDI-II) to measure the depression severity and used a cut-off value (BDI-II score $\geq$ 14) to decide whether a participant was depressed or not. In our study, depression status is based on clinical assessment, and hence is more reliable than that based on self-reports in [58]. Secondly, we extracted a comprehensive set of features from SMS and phone call logs, while the study in [58] only used a few basic statistics. Last, as mentioned above, in [58], SMS and phone call statistics were used together with several other types of data (including average daily usage of mobile phones, amount of time spent on web, social media apps, and entertainment apps, age, gender) to predict depression, while our study focuses on the effectiveness of depression prediction using a single type data (SMS or phone call).

Using other types of sensing data for mental health applications. Many existing studies have used smartphone location and activity data for stress and depression screening. Wang et al. [81] found significant correlation between the behavioral features (in terms of conversation duration, number of locations visited, sleep) and depressive mood in college students. Saeb et al. [61] found significant correlation between the phone usage and mobility patterns with respect to the self-reported

depressive scores. Canzian and Musolesi [13] studied the relationship between the mobility patterns and depression, and found that individualized machine learning models outperformed general models. Farhan et al. [25] found that location and activity related features extracted from the smartphone sensing data can predict depression with good accuracy. Yue et al. [88] investigated fusing two types of location data, GPS and WiFi association data, both collected locally on phones, for more complete location information for improved depression detection. Lu et al. [49] developed a heterogeneous multi-task learning approach for analyzing sensor data collected over multiple smartphone platforms. All the above studies use location and/or activity data, while our study investigates using social interaction data for depression prediction.

4.3 Data Collection

We collected data from a two-phase study at the University of Connecticut. Phase I study was from October 2015 to May 2016; Phase II study was from February 2017 to December 2017. The participants were full-time students of the university, aged 18-25, using either iPhones or Android phones. We were only able to collect SMS and phone call logs on Android phones (the restrictions of iOS prevented us from collecting such data on iPhones). Therefore, we only consider a subset of participants, i.e., the Android users, in this study. We recruited a total of 59 Android users (25 and 34 users in the two phases, respectively). Based on clinician diagnosis, 18 users were depressed and 41 users were non-depressed. We next briefly describe the two types of data that are used in this work: in-phone communication data (i.e., SMS and phone call logs) and clinical assessment.

4.3.1 SMS and Phone Call Logs

We used an app that we developed, called *LifeRhythm* [25], to log SMS and phone call records on Android phones. Specifically, the app queried the SMS and call logs once a day and recorded all the data corresponding to that day. The content of the SMS messages or phone calls was never recorded. To ensure the privacy of the participants, we assigned a random ID to each participant, which was used to identify the participants. The smartphone sensing data collected by the app was encrypted and sent to a secure server when the phone was connected to a WiFi network. After the data reached the server, the server decrypted the data, hashed the phone numbers of the contacts in the records to preserve user privacy, and then stored the data in a database. Only statistical information was used in the analysis (see Section 4.4).

SMS data. Each SMS record corresponds to an SMS messaging event, represented as a tuple $(s_i, h_i, t_i, w_i, c_i)$, where i is the row index of the event, s_i is the sense time of the event, h_i is the hashed phone number that the participant was communicating with, t_i is the type of the message (i.e., incoming or outgoing), w_i is the number of words in the message, and c_i is the number of characters in the message.

Phone call records. Each phone call record corresponds to a phone call event, represented as a tuple (s_i, h_i, t_i, d_i), where, similar as an SMS event, i is the row index of the event, s_i is the sense time of the event, h_i is the hashed phone number that the participant was communicating with, t_i is the type of the call (i.e., incoming, outgoing or missed call), and d_i is the duration of the call.

4.3.2 Clinical Assessment

Every participant filled in a self-report questionnaire at the beginning of the study. The questionnaire used in Phase I study was Patient Health Questionnaire (PHQ-9) [44]. In Phase II study, it was changed to Quick Inventory of Depressive Symptomatology (QIDS) [60] because QIDS provided more detailed information on patient symptoms than PHQ-9. A participant was screened initially by our study clinician. Using an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and self-report PHQ-9/QIDS evaluation, the clinician classified individuals as either depressed or non-depressed during the initial screening. All participants filled in PHQ-9/QIDS at a regular basis on their phones while in the study (a notification was sent to their phones at the due date, every 14 days for PHQ-9 and 7 days for QIDS). A participant with a diagnosis of depression must participate in treatment to remain in the study. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician) to confirm their self-reported PHQ-9/QIDS scores with their verbal report during the meetings.

4.4 Feature Extraction

We extract a comprehensive set of features from SMS data and phone call logs. These features characterize the quantity, timing and variability of incoming and outgoing messages or calls. All the features represent statistical information, not related to the content of the messages or phone calls (the content was never captured).

4.4.1 Feature Extraction for SMS Data

We extracted the following 29 features from the SMS data. These features are broadly in two categories that are related to (i) the basic statistics (quantity, timing and length) of the messages, and (ii) variability of the messages (in standard deviation and entropy, and unique contacts). We represent the features for incoming and outgoing messages separately since they represent passive and user initiated activities, respectively. In addition, we include basic statistics for different times of the day since depressed and non-depressed users may exhibit different temporal usage patterns. Specifically, we consider three time periods, morning (from 6 am to 12 pm), afternoon (from 12 pm to 6 pm) and evening (from 6 pm to 6 am). The length of a message is represented as the total number of characters in the message (we did not find significant differences between representing the length in characters and words). All the features are defined for a time interval of n days, $n \geq 1$.

# of incoming messages. This feature represents the total number of messages received.

# of outgoing messages. This feature represents the total number of messages sent.

# of incoming messages (morning, afternoon, evening). These three features represent the numbers of messages received during different time periods of a day, i.e., morning, afternoon and evening, respectively.

# of outgoing messages (morning, afternoon, evening). These three features represent the numbers of messages sent in different time periods of a day, i.e., morning, afternoon and evening, respectively.

Average length of incoming messages. This feature represents the average length

of the received messages. It is calculated by dividing the total length of the received messages by the total number of received messages.

Average length of outgoing messages. This feature represents the average length of the sent messages. It is calculated by dividing the total length of the sent messages by the total number of sent messages.

Average length of incoming messages (morning, afternoon, evening). These three features are the average lengths of the messages received in different time periods of a day, i.e., morning, afternoon and evening, respectively.

Average length of outgoing messages (morning, afternoon, evening). These three features are the average lengths of the messages sent in different time periods of a day, i.e., morning, afternoon and evening, respectively.

Standard deviation of incoming message length. This feature calculates the standard deviation of the lengths of the received messages.

Standard deviation of outgoing message length. This feature calculates the standard deviation of the lengths of the sent messages.

# of unique contacts. This feature represents the total number of unique contacts from whom a user received messages or to whom a user sent messages.

# of unique contacts (incoming). This feature represents the number of unique contacts from whom a user received messages.

# of unique contacts (outgoing). This feature represents the number of unique contacts to whom a user sent messages.

Entropy of # of incoming messages. This feature measures the variability of the number of messages that a user received from unique contacts (only including the contacts from whom the user received messages). Let p_i be the percentage of the

number of messages that a user receives from contact i. Then this feature is defined as $\sum_i -p_i \log p_i$.

Normalized entropy of # of incoming messages. Let N_r be the number of unique contacts from whom the user received messages. Then we further define normalized entropy for the above feature, which is the entropy value normalized by $\log N_r$ so that the it is invariant to N_r and depends solely on the distribution of the number of received messages.

Entropy of length of incoming messages. We further define entropy features related to the lengths of the messages that a user received from unique contacts. It differs from the previous feature in that p_i is the ratio of the total length of messages that the user received from contact i over the total length of messages that the user received.

Normalized entropy of length of incoming messages. This is the normalized entropy of the previous feature.

Entropy and normalized entropy of outgoing messages. Similar as received messages, we define two entropy features related to sent messages, in terms of the number and the length of the messages, respectively. We further define normalized entropy for each of them.

4.4.2 Feature Extraction for Phone Call Logs

We extracted 30 features from the phone call logs. Of them, 29 features are similar as those defined for SMS messages: the features for incoming and outgoing calls are defined separately; the "length" of a call is the duration of the call in minutes. One additional feature that is defined for phone calls, while not defined for SMS, is the

TABLE 4.1: Characteristics of SMS data for depressed and non-depressed participants.

Feature	Depressed		Non-depressed		p-value
	Mean	Stdev	Mean	Stdev	
Daily # of incoming messages	21.36	20.84	12.25	5.60	0.19
Daily # of outgoing messages	24.22	26.56	11.64	6.50	0.18
Daily # of incoming messages (morning)	4.22	4.16	2.45	1.10	0.20
Daily # of incoming messages (afternoon)	7.60	6.47	4.48	2.00	0.17
Daily # of incoming messages (evening)	9.55	12.57	5.32	3.01	0.25
Daily # of outgoing messages (morning)	4.21	5.20	2.27	1.72	0.23
Daily # of outgoing messages (afternoon)	8.61	8.58	4.11	1.99	0.15
Daily # of outgoing messages (evening)	11.40	14.72	5.25	3.55	0.20
Daily avg. incoming message length	47.42	15.28	48.94	17.63	0.43
Daily avg. outgoing message length	38.98	12.60	46.21	15.31	0.18
Daily avg. incoming message length (morning)	38.00	19.31	37.15	25.79	0.47
Daily avg. incoming message length (afternoon)	38.11	8.54	36.68	15.15	0.41
Daily avg. incoming message length (evening)	29.92	5.78	34.18	10.09	0.16
Daily avg. outgoing message length (morning)	17.28	11.68	21.53	9.71	0.25
Daily avg. outgoing message length (afternoon)	28.72	11.22	35.25	12.92	0.17
Daily avg. outgoing message length (evening)	29.94	14.13	30.86	10.40	0.45
Standard deviation of incoming message lengths (per day)	32.87	7.66	35.00	13.86	0.35
Standard deviation of outgoing message lengths (per day)	24.64	8.75	31.17	10.07	0.11
Daily # of unique contacts	3.88	2.40	2.86	0.86	0.20
Daily # of unique contacts (incoming)	2.81	0.76	2.63	0.84	0.35
Daily # of unique contacts (outgoing)	2.43	0.96	2.38	0.81	0.46
Entropy of daily # of incoming messages	0.63	0.13	0.63	0.23	0.47
Normalized entropy of daily # of incoming messages	0.19	0.05	0.19	0.05	0.50
Entropy of daily length of incoming messages	0.64	0.16	0.59	0.23	0.32
Normalized entropy of daily length of incoming messages	0.19	0.04	0.18	0.04	0.31
Entropy of daily # of outgoing messages	0.24	0.15	0.47	0.24	0.03
Normalized entropy of daily # of outgoing messages	0.09	0.05	0.14	0.05	0.04
Entropy of daily length of outgoing messages	0.22	0.14	0.42	0.24	0.03
Normalized entropy of daily length of outgoing messages	0.08	0.05	0.13	0.05	0.04

number of missed calls.

4.5 Characteristics of SMS and Phone Call Logs

In this section, we characterize the various SMS and phone call features for depressed and non-depressed participants. Of the 59 participants, we were only able to collect SMS logs from 15 participants (5 depressed and 10 non-depressed) during the study period, maybe because many participants used other messaging apps (e.g., What-

sApp, Snapchat) instead of SMS. For phone calls, we were able to collect data from 46 participants (16 depressed and 30 non-depressed), maybe because most of the participants still used the standard phone-call services built in phones. Overall, for SMS, we collected a total of 1139 days of data from the 15 users (476 from depressed users and 663 from non-depressed users). For phone call logs, we collected a total of 2538 days of data from the 46 users (982 from depressed users and 1556 from non-depressed users).

In the following, we consider daily feature values, i.e., the value of a feature was obtained using one day's data. For one user, we obtained the feature value for each day and then obtained the average value over the days with data. For each feature, we present the mean and standard deviation across the users for depressed and non-depressed populations separately. In addition, for each feature, we performed a two-sample independent one-tailed t-test on the difference of the mean values of the depressed and non-depressed users (i.e., the null hypothesis is that the means of the two populations are the same, while the alternative hypothesis is that the attribute of the depressed population is larger or smaller than that of the non-depressed population, where larger or smaller is selected based on individual attributes), and obtained the p-value. The results for the SMS and phone call logs are shown in Tables 4.1 and 4.2, respectively. We next summarize the main results.

Characteristics of SMS data. We observe from Table 4.1 that five features have significant p-values: standard deviation of outgoing message length (with p-value of 0.11), entropy and normalized entropy of number of outgoing messages (with p-values below 0.05), and entropy and normalized entropy of length of outgoing messages (with p-values below 0.05). All of these five features are related to the variability of outgoing messages, and the variability for depressed participants is lower than that

of non-depressed participants. The above observations indicate that characteristics of outgoing messages, which were initiated by the users, are more distinguishing between depressed and non-depressed participants than incoming messages. Considering outgoing message length, it appears that depressed participants tend to send shorter messages, with a lower standard deviation of these messages. In addition, they tend to send most messages to a lower number of contacts, leading to lower entropy values than non-depressed participants (although for depressed participants, the average number of unique contacts for outgoing messages is not necessarily lower than that for non-depressed participants).

Characteristics of phone call logs. In Table 4.2, 10 features have p-values at significant level of 0.10: number of incoming calls, number of outgoing calls, number of outgoing calls (morning), number of outgoing calls (afternoon), duration of outgoing calls (afternoon), standard deviation of outgoing call duration, number of unique contacts, number of unique contacts for outgoing calls, and entropy and normalized entropy of outgoing calls. Except for two features (number of incoming calls and number of unique contacts), the rest of the features above are for outgoing calls, again indicating that actions initiated by the users are more differentiating between depressed and non-depressed populations than phone calls received by the users. Interestingly, we observe that depressed users have more outgoing calls, longer outgoing calls, higher standard deviation in outgoing calls, more unique contacts for outgoing calls and larger entropy for outgoing calls. For the two features not related to outgoing calls, we also observe that depressed participants have more incoming calls and unique contacts. Overall, it appears that depressed participants spent more time on phone call related activities. This might be because social interaction of depressed population is more confined towards indirect mode of communication, i.e., over the

TABLE 4.2: Characteristics of phone call logs for depressed and non-depressed participants. The duration of a call is in minutes.

Feature	**Depressed**		**Non-depressed**		**p-value**
	Mean	Stdev	Mean	Stdev	
Daily # of incoming calls	2.02	0.91	1.65	0.36	0.07
Daily # of outgoing calls	3.28	1.91	2.51	0.88	0.07
Daily # of missed calls	0.20	0.21	0.17	0.17	0.29
Daily # of incoming calls (morning)	0.36	0.26	0.26	0.14	0.08
Daily # of incoming calls (afternoon)	0.86	0.43	0.75	0.21	0.16
Daily # of incoming calls (evening)	0.79	0.47	0.65	0.25	0.14
Daily # of outgoing calls (morning)	0.61	0.57	0.38	0.20	0.07
Daily # of outgoing calls (afternoon)	1.51	0.92	1.12	0.54	0.07
Daily # of outgoing calls (evening)	1.16	0.67	1.00	0.47	0.21
Daily avg. incoming call duration	5.74	8.17	4.47	3.20	0.28
Daily avg. outgoing call duration	4.30	5.88	3.36	2.21	0.27
Daily avg. incoming call duration (morning)	1.05	1.28	0.62	0.62	0.11
Daily avg. incoming call duration (afternoon)	2.26	1.06	1.87	1.23	0.13
Daily avg. incoming call duration (evening)	4.02	8.04	2.75	2.66	0.27
Daily avg. outgoing call duration (morning)	0.73	0.66	0.81	1.12	0.38
Daily avg. outgoing call duration (afternoon)	2.63	2.75	1.42	0.76	0.05
Daily avg. outgoing call duration (evening)	2.97	4.08	2.00	1.83	0.19
Standard deviation of incoming call duration (per day)	1.80	2.31	1.05	0.91	0.11
Standard deviation of outgoing call duration (per day)	2.24	1.64	1.50	1.05	0.06
Daily # of unique contacts	2.90	0.85	2.60	0.47	0.10
Daily # of unique contacts (incoming)	1.51	0.38	1.40	0.20	0.15
Daily # of unique contacts (outgoing)	2.28	0.89	1.94	0.51	0.08
Entropy of daily # of incoming calls	0.29	0.18	0.24	0.11	0.15
Normalized entropy of daily # of incoming calls	0.12	0.06	0.10	0.04	0.13
Entropy of daily duration of incoming calls	0.19	0.13	0.15	0.08	0.13
Normalized entropy of daily duration of incoming calls	0.08	0.05	0.06	0.03	0.13
Entropy of daily # of outgoing calls	0.37	0.28	0.27	0.16	0.09
Normalized entropy of daily # of outgoing calls	0.11	0.05	0.09	0.04	0.07
Entropy of daily duration of outgoing calls	0.22	0.19	0.16	0.12	0.12
Normalized entropy of daily duration of outgoing calls	0.07	0.04	0.05	0.03	0.14

phone calls as compared to more direct mode of in-person communication [42].

Summary. Our observations of the characteristics of SMS and phone calls differ from those in [58], which found that participants with depression (i) sent more text messages, and (ii) made and received fewer calls, with shorter call duration (for both incoming and outgoing calls). For SMS, while our data shows that on average the depressed participants indeed sent more messages than non-depressed participants, we did not find substantial evidence to reject the null hypothesis that their mean values are the same (see Table 4.1). Instead, our main finding is that various variability attributes of outgoing messages can differentiate depressed and non-depressed participants more significantly, which were not considered in [58]. For phone calls, we found that depressed participants actually made and received more phone calls, the opposite of what observed in [58]. In addition, we did not find that phone call duration for depressed participants to be shorter than that of non-depressed participants; in fact, there is strong evidence that depressed participants had longer outgoing calls in the afternoon than non-depressed participants. We again find that various variability attributes of the outgoing calls can differentiate depressed and non-depressed participants, which were not considered in [58]. For the features considered in both [58] and our study, the different observations might be due to different demographics: participants in our study were all college students, while the participants in the study in [58] were from much more diverse backgrounds and age groups. In addition, the depression status in our study was based on clinician assessment, while was based on self-report scores in [58].

4.6 Depression Prediction

In this section, we explore using SMS and phone call features for depression prediction. We use these two types of data separately since, as mentioned earlier, we may not be able to obtain both types of data from a user. For instance, a user may choose to use other messaging app (instead of SMS), while use the built-in calls directly on a phone. Our focus is to quantify the effectiveness of depression screening using only one type of data; when both types of data are available, the classification accuracy can be further improved. For the classification using SMS and call data, we used fractional values for number of incoming and outgoing messages/calls in different times of the day i.e. morning, afternoon and evening. These values are normalized by the total number of messages received so that their sum is 1. The normalization allows the features to be more resilient to missing data. For each user, we consider a moving window of n days and make a classification on a daily basis. Specifically, for each day t, we consider the data collected during the past n days, i.e., $[t-n+1, t]$, to classify the depression status (i.e., whether the user is depressed or not), as illustrated in Fig. 4.1. The results in the rest of the chapter were obtained for $n = 14$ days; we also explored using $n = 7$ days, and the results were not as good as using 14 days of data.

We observed missing data (i.e., no data was collected) on some days, which may be due to various reasons, e.g., failed data collection, malfunction of the phone, or no activity from a user. Fig. 4.2a plots cumulative distribution function (CDF) of the number of consecutive days with missing data for the SMS dataset; the results for three cases (all the users, the depressed users and non-depressed users) are shown in the figure. We see that for approximately 87% of the cases, the number of consecutive days with missing data is no more than 3 days. Fig. 4.2b plots the corresponding

results for phone call logs, showing that for approximately 83% of the cases, the number of consecutive days with missing data is no more than 3 days.

Based on the above observations, we consider three scenarios in the following. In an interval of n days, let k be the maximum number of consecutive days with missing data that is allowed in the window. The three scenarios we consider correspond to $k = 0$, $k = 1$ and $k = 3$. That is, in the first scenario, we consider an interval only if there is no missing data in that interval; in the latter two scenarios, the number of consecutive days with missing data cannot exceed 1 and 3, respectively. The third scenario is the least strict in the amount of missing data, and covers most of the samples based on the observations in Figures 4.2a and 4.2b.

As mentioned earlier, we collected SMS logs from 15 participants (5 depressed and 10 non-depressed) and phone call logs from 46 participants (16 depressed and 30 non-depressed). For SMS, when $k = 0$, a total of 12 users (4 depressed, 8 non-depressed) had valid samples, i.e., having at least one interval of $n = 14$ days with no missing data. When $k = 1$ and 3, the corresponding numbers of users were 14 (5 depressed, 9 non-depressed) and 17 (5 depressed, 12 non-depressed), respectively. Figures 4.3a to 4.3c plot the number of samples contributed by each user who had at least one valid sample for the SMS dataset, corresponding to $k = 0$, 1, and 3, respectively. For phone calls, when $k = 0$, a total of 30 users (13 depressed, 17 non-depressed) had valid samples; when $k = 1$ and 3, the corresponding numbers of users were 34 (15 depressed, 19 non-depressed) and 46 (16 depressed, 30 non-depressed), respectively. Figures 4.3d to 4.3f plot the per user sample distribution of the three scenarios for the phone call dataset.

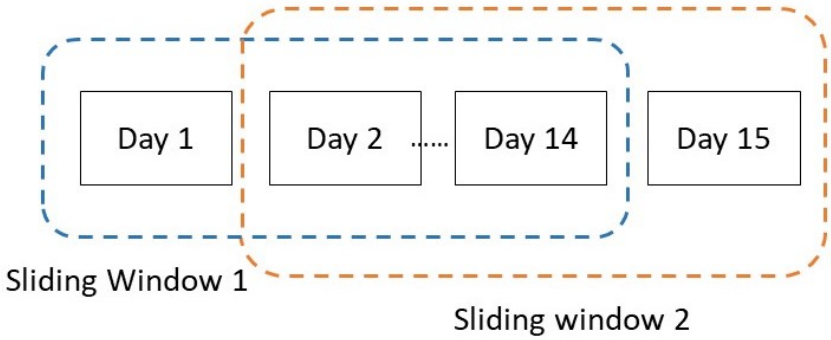
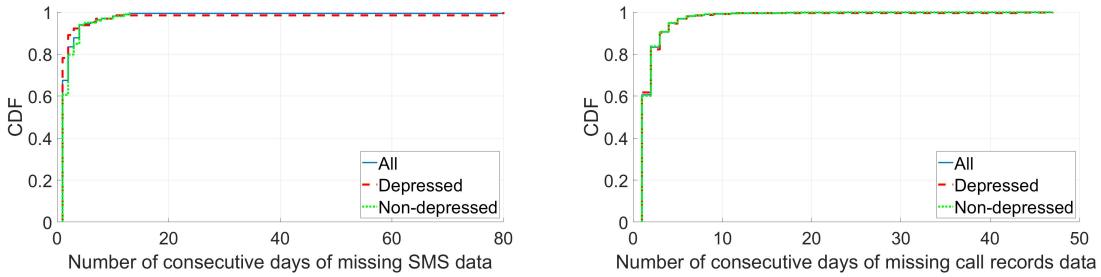


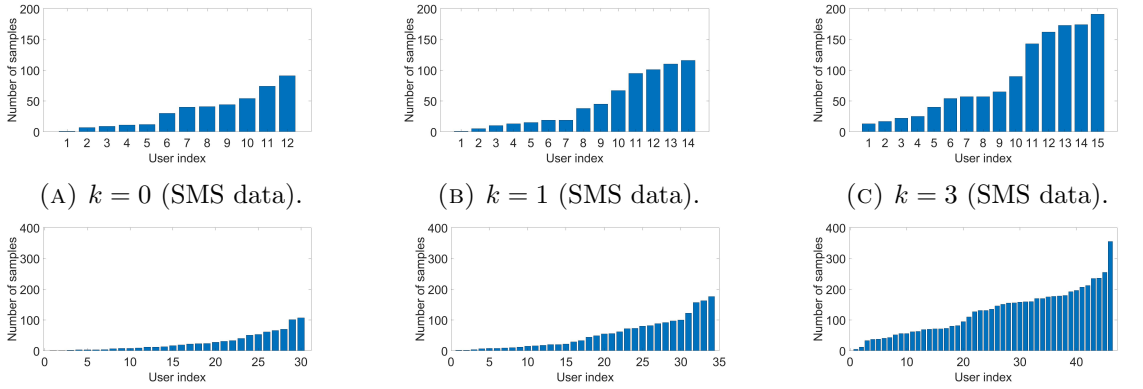
FIGURE 4.1: Illustration of using data collected in a sliding window of n days for depression prediction. Here $n = 14$.



(A) SMS logs.

(B) Phase I daytime monitoring

FIGURE 4.2: Number of consecutive days with no data sample.



(A) $k = 0$ (SMS data).

(B) $k = 1$ (SMS data).

(C) $k = 3$ (SMS data).

(D) $k = 0$ (phone call data).

(E) $k = 1$ (phone call data).

(F) $k = 3$ (phone call data).

FIGURE 4.3: Number of valid samples from each user for SMS and phone call datasets, where $k = 0$, 1, or 3.

4.6.1 Classification Methodology

We explored three classification algorithms: Support Vector Machine (SVM) with radial basis function (RBF) kernel [14], random forest classifier [10] and XGBoost [15] for depression prediction. The classification was done for an interval of $n = 14$ days using the various features derived in the interval as input to the classification algorithms. The clinical ground truth served as the label for depression status (i.e., whether a user is depressed or not). We used leave-one-user-out cross validation procedure (i.e., no data from one user was used in both training and testing to avoid overfitting).

Among the three classification algorithms, we found that XGBoost led to the best prediction results. In the following, in the interest of space, we only present the results using XGBoost. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. We further used Extra-Trees Classifier (ETC) [30, 48] to determine the importance of the features. With ETC, random trees are constructed from subsamples of the training dataset. For each feature under consideration, a random value that is selected from the feature's empirical range is selected for the split. ETC returns the importance scores for all the features; the higher the values is, the better the feature is. When using XGBoost, we chose the top m features (based on the ranking from ETC), and varied m from 1 to the total number of features. The set of m features in combination with parameter tuning of XGBoost (see below) that provided the highest F_1 score was chosen as the best set of features. The F_1 score, defined as $2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$, is a weighted average of the precision and recall. It ranges from 0 to 1, and the higher,

TABLE 4.3: Depression prediction results using XGBoost.

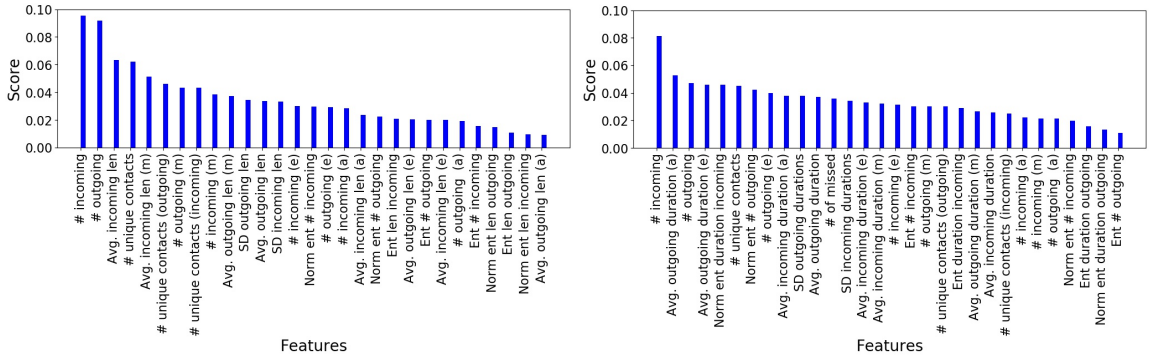
	Scenario	F_1 Score	Precision	Recall	Specificity	# of features selected
SMS logs	$k = 0$	0.78	0.69	0.89	0.56	10
	$k = 1$	0.80	0.82	0.79	0.82	12
	$k = 3$	0.69	0.66	0.72	0.67	13
Phone call logs	$k = 0$	0.78	0.66	0.89	0.56	14
	$k = 1$	0.71	0.60	0.87	0.58	17
	$k = 3$	0.73	0.69	0.78	0.82	5

the better.

For XGBoost, we used parameter tuning along with leave one-user-out cross validation. Specifically, we varied the following parameters and chose the values that gave the best validation F_1 score: the maximum depth of a tree was varied from 3 to 10, the minimum child weight (i.e., the minimum sum of weights of all observations required in a child of a tree, which was used to control over-fitting) was varied from 1 to 6, the fraction of observations to be randomly sampled for each tree and the fraction of features to be randomly sampled for each tree were both varied from 0 to 1, and the gamma value (i.e., the minimum loss reduction required to make a further partition on a leaf node of a tree) was varied from 0 to 1. Throughout, we used a learning rate of 0.01.

4.6.2 Depression Prediction Using SMS Data

The top half of Table 4.3 presents the classification results using SMS data in the three scenarios of $k = 0$, 1 and 3. We see similar F_1 scores, 0.80 and 0.78, when $k = 0$ and $k = 1$; when allowing more consecutive days with missing data (i.e., when $k = 3$), the F_1 score is lower. Overall, the F_1 scores are comparable to those obtained using location and activity sensors [61, 13, 89, 25]. For each scenario, out of the 29



(A) SMS data. (B) Phone call logs.

FIGURE 4.4: Importance scores of the features calculated by ETC method when $k = 0$, where '(m)', '(a)' and '(e)' represent the time periods of morning, afternoon and evening, respectively.

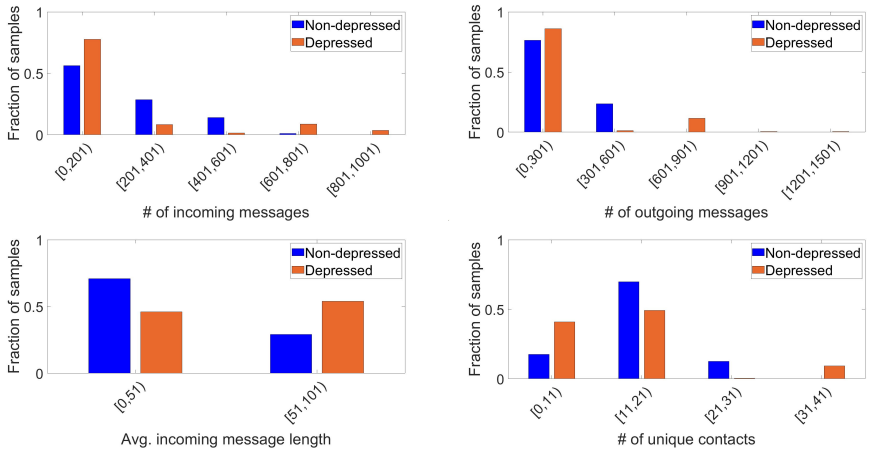


FIGURE 4.5: Histograms of the top 4 selected features for SMS data, $k = 0$.

features, the number of selected features was from 10 to 13.

In the interest of space, we only present the selected features for the scenario of $k = 0$. Fig. 4.4a plots the importance scores of the features calculated by the ETC method (higher score indicates better feature). From Table 4.3, the top 10 features were selected when $k = 0$. We see that both types of features, i.e., features that represent the overall statistical information and the features for particular time periods, were selected. The former type of features includes total number of incoming messages, total number of outgoing messages, average length of incoming messages, number of unique contacts considering all messages, number of unique contacts considering only outgoing messages, and number of unique contacts considering only incoming messages; the latter type of features include average length of incoming messages in the morning, number of incoming messages in the morning, number of outgoing messages in the morning, and the average length of outgoing messages in the morning.

Fig. 4.5(a)-(d) plots the histogram of the top 4 selected features for samples from depressed and non-depressed separately. We see from Fig. 4.5(a) that the number of incoming messages to the depressed participants tend to be low; on the other hand, in a substantial fraction of the instances, depressed participants have a large number of incoming messages. Similar observations hold for the number of outgoing messages as shown in Fig. 4.5(b). In terms of average length of incoming messages, a higher fraction of instances from depressed participants have longer messages than that from non-depressed participants (see Fig. 4.5 (c)). Last, while the number of unique contacts is low for most instances from the depressed participants, a noticeable fraction of the samples from the depressed participants have large number of unique contacts (see Fig. 4.5 (d)).

4.6.3 Depression Prediction Using Phone Call Logs

The bottom half of Table 4.3 presents the classification results when using phone call logs. The F_1 score ranges from 0.71 to 0.78 across various scenarios, comparable to the F_1 scores obtained using location and activity sensors [61, 13, 89, 25]. The highest F_1 score was obtained when $k = 0$. The F_1 scores when $k = 1$ and $k = 3$ were comparable. Out of the total 30 features, the number of selected features were from 5 to 17.

Again, in the interest of space, we only present the selected features when $k = 0$. Fig. 4.4b plots the importance scores of all 30 features calculated by ETC. The top 14 features were selected, including features on the overall phone usage attributes (e.g., number and length of calls), variability attributes (e.g., normalized entropy of duration of incoming calls), and the information for particular time periods (average duration of outgoing calls in the afternoon and in the evening). Fig. 4.6 plots the histogram of the top 4 features for samples from the depressed and non-depressed participants. Fig. 4.6(a) shows that, for the number of incoming calls, a larger fraction of samples from the depressed participants have higher number of calls than that from the non-depressed participants. The same observation holds for the number of outgoing calls as well (see Fig. 4.6(b)). For all outgoing calls in the afternoon and evening (see Fig. 4.6(b) and (d)), we observe that majority of the samples from non-depressed population were short (in a few minutes). For depressed population, while majority of the samples also have short duration, some samples are of significantly higher duration.

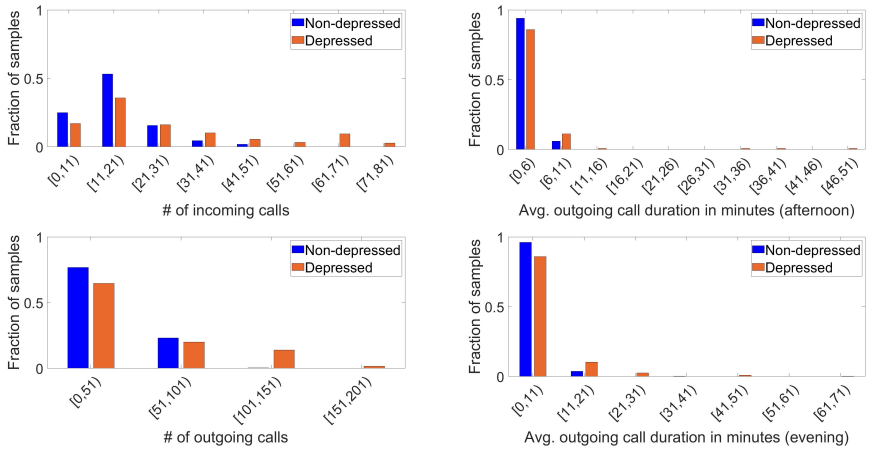


FIGURE 4.6: Histogram of the top 4 selected features for phone call logs, $k = 0$.

4.7 Conclusion and Future Work

In this chapter, we have investigated using social interaction data, specifically SMS and phone call logs, passively collected on smartphones for depression prediction. We extracted a comprehensive set of features from SMS and phone call logs, and compared the features of depressed and non-depressed participants. We have constructed a family of machine learning models using these features to predict depression and the best models (using XGBoost) lead to F_1 score up to 0.80. Overall, our results demonstrate that the SMS and phone call logs alone can already provide useful insights into behavioral patterns of individuals, which can be used to effectively predict depression.

Limitations of the study and future work. All participants in our study were college students. Future work includes examining other demographics. Our sample size, particularly for SMS, is small since many users were not using SMS actively on their phones. On the other hand, we believe a substantial number of users will remain using built-in phones call due to their convenience, and a natural

future direction is validation of the results using phone call logs from a larger number of participants. Another interesting future direction is to explore using other social interaction data, e.g., email logs and social media data, collected on smartphones for depression screening.

Chapter 5

Conclusion

With the expanding research in the area of mobile computing for mental health, a number of studies have contributed some very innovative solutions for solving the depression as a public health problem. Each individual face this issue at some point in life, which often leads to serious consequences on physical and mental health and often leading to suicide. There is indeed a very genuine need to detect depression at an early stage and prevent it from getting worse.

In the first part of the dissertation, we present a novel approach that uses metadata from WiFi infrastructure for automatic depression screening. We extracted features at both AP and building level and calculated their correlation with the self-report scores. We found that, our analysis over the two datasets from Phase I and Phase II can predict depression status effectively. We also found that using the behavioral features from the WiFi association records, we could construct multi-feature regression models to predict PHQ-9 and QIDS scores. The prediction results are comparable to those obtained using the data collected by instrumenting the individual

phones.

In the second part of this dissertation, we have investigated the feasibility of using the smartphone data i.e. smartphone sensing data and WiFi infrastructure meta-data to predict individual depressive symptoms. Our results indicate that all major categories of both behavioral and cognitive symptoms can be predicted fairly accurately using smartphone data. We also found that the finer level of some sleep related and psychomotor depressive symptoms could also be predicted accurately using smartphone data. Our study makes an important step forward over existing studies in demonstrating that using passively collected smartphone data is a promising direction in automatically keeping track of depressive symptoms.

In the third part of the dissertation, we investigated the the possibility of using social interaction data particularly SMS and phone call records for predicting the depression status. We calculated features across different times of the days i.e. morning, afternoon and evening and also considered different time frames (overlapping days) for data analysis. Our results demonstrated that the SMS and phone call records can be effectively used for predicting the depression accurately. We also found that the individuals with depression tend to spend more time on phone calls. On the contrary, they use SMS less frequently. Our results are comparable to the results obtained from studies that have only used location data for depression prediction. Our study shows that consideration of social interaction behavior is as crucial as behavioral patterns derived using location visits of individuals.

5.1 Insights

This dissertation presented different novel approaches related to automatic depression screening. We took the research platform related to smartphone sensing for mental health to one step ahead by not just exploring the usage of smartphone sensing data but also instrumenting the meta-data passively collected from WiFi infrastructure. We could demonstrate that such approach can achieve comparable performance for depression screening as the approach based on instrumenting smartphones. We envision this type of depression diagnostic tool could be deployed at both population level like a campus like setting, for example, a university, a military camp, workplace and others. It could also be deployed at individual level which can be immensely useful for clinicians to make effective treatment decisions. We considered two scenarios for data analysis i.e. 24-hour monitoring that is applicable to students living on campus and daytime (8 am to 6 pm) monitoring that is applicable to commuting scenario. We considered 3 levels of analysis namely at AP level, building level and enhanced building level for both the scenarios. Significant negative correlation between entropy and PHQ-9 scores indicates that participants tend to spend more time in few locations. Positive correlation between time at home and PHQ-9 scores indicates that participants spend more time at home. For multi-linear regression, we found that the non-linear model has a stronger correlation with the self-report scores than the linear model. For classification results to predict depression status, we found better results using building-level features than AP-level features. For AP level analysis, we found the classification results for daytime monitoring are better than 24-hour monitoring. The top features selected by the feature selection methods included features like average duration spent in library and sports buildings, number of days of visiting

library building. This reflects that adding building semantics can lead to even better results and prove more intuitive for depression prediction. Our results indicate that using features leads to better results than using self-report scores (F_1 score 0.68).

Next, this dissertation presented another novel approach for predicting individual depressive symptoms. We used smartphone sensing data and WiFi infrastructure meta-data for predicting depressive symptoms. We found that using behavioral data i.e. location visits of users, both behavioral and cognitive symptoms can be predicted accurately. Using smartphone sensing data for both Phase I and Phase II, we found that symptoms like appetite, interest and feeling depressed were predicted accurately for depressed participants. For all the participants, concentration, energy, feeling-depressed, interest, self-criticism, and sleep were predicted accurately in various settings. When using the WiFi infrastructure meta-data for predicting depressive symptoms, we observed that all the seven symptoms that were analyzed were predicted accurately for both Phase I and Phase II.

In the last part of this dissertation, we analyzed SMS and phone call data for depression prediction. We used multiple classifiers i.e. SVM with SVM-RFE, random forest classifier and XGBoost with Extra-Trees classifier feature selection methods to select a subset of features to improve the performance. We found that the highest F_1 is obtained by using XGBoost classifier. Our results show that the depressed individuals send and receive less messages than the overall average. Also, the depressed population receives longer messages than the overall average. For phone call data, we observe that the depressed population spends more time on phone calls (for both incoming and outgoing).

To summarize this dissertation, we conclude that the approaches presented in this dissertation effectively uses the smartphone data i.e. smartphone sensing data

and WiFi infrastructure meta-data for automatic depression screening that could be deployed at both population and individual levels. This would serve as a strong platform to keep tabs on the mental health of individuals in campus setting as well as resulting in effective treatment decisions by the clinicians.

5.2 Future Work

Despite the significant contributions of our research work, there are a number of ways in which our work can be improved and extended. The following are some directions that future studies could explore:

- **Other campus settings.** Our study is focused on the college-age students from the campus at the University of Connecticut. It would be intuitive to explore our approach in other university campuses, and in other settings (e.g., workplace, military base).

- **Analysis in other demographics and gender specific studies.** It would be interesting to explore the approaches presented in this dissertation in other demographic groups. Similarly, by analyzing and organizing the results specific to gender would also provide useful insights into behavioral patterns.

- **Modeling human behavior.** Our data preprocessing methodologies considered several aspects of anomalies due to the stochastic nature of human behavior while calculating the features. Some of them included time specific thresholds to decide if a participant indeed spent time in a building or just passed by, considering different times of the day for analysis, identifying a home cluster if a user spent significant amount of time in a place between 12 am to 6 am. We

also considered different thresholds (decided empirically) to include days with data or exclude consecutive days with no data sample at all. However, there are a lot of other behavioral lifestyles that should also be taken into consideration like bedtime schedules, daily routine, work life if considering a workplace population etc.). This would help in constructing more precise and effective models for depression prediction.

- **Using other smartphone sensing data.** Our studies used a wide array of sensors like GPS, WiFi, activity, SMS, phone call records for depression prediction. However, usage of other sensor information like screen, light, email history etc. would be valuable as well.

- **Handle missing data.** We have organized our analysis in a reasonably effective way that handles missing data to quite some extent. However, better techniques could be investigated to further improve the data quality and handle the missing data.

- **Feature extraction.** We have calculated a wide variety of location, activity and social interaction based features that cover different flavors of human behavior. However, more meaningful features could be included for more effective prediction results.

- **Feature selection methods.** We have used SVM-RFE for selecting subset of top features that improve the predictor performance. But, there is a limitation to this approach as we have calculated the overall rank of features by averaging the individual ranks returned by SVM-RFE for different *cost* and γ values. In the future, we plan to explore more efficient feature selection methods that are

independent of parameters.

- **Improve F_1 scores.** In the future, we plan to explore more effective data preprocessing and classification methods that would lead to better prediction results with higher F_1 scores.

- **Combine different sources for depression prediction.** In our current works, we have explored the feasibility of using different types of smartphone sensing data including GPS, WiFi, SMS and call logs for depression prediction. While we have used each of these sources independently, it is still unclear, which sensing information is more effective in predicting depression and depressive symptoms. In the future, we plan to investigate the feasibility of combining different sensing information together for effective depression prediction. There are many challenges associated when combining them together like handling missing data, mapping various timestamps for data points etc.

- **Cryptography techniques.** Our works ensured user privacy by standard cryptographic techniques. However, there is score to develop more systematic techniques to ensure user privacy and data security.

- **Designing a clinician acceptable diagnostic system.** A systematically designed tool that could be used by the clinicians and help them understand the results returned by the various pre-trained machine learning models is required. Eventually, the main aim is to develop a system that would provide real time data analytics for depression prediction which in turn, could help clinicians observe the course of changes in the behavior of patients to identify the triggers and make effective treatment decisions.

Bibliography

[1] "Centers for Disease Control and Prevention. National Center for Injury Prevention and Control." 2010, `http:www.cdc.gov/ncipc/wisqars`.

[2] "Results from the 2017 national survey on drug use and health: Detailed tables," https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.htm#tab8-56A, 2017.

[3] C. Aguilar-Melchor and P. Gaborit, "A lattice-based computationally-efficient private information retrieval protocol," in *Proc. of Western European Workshop on Research in Cryptology (WEWoRC)*, July 2007.

[4] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting mobile 3g using wifi," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 209–222.

[5] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2009, pp. 280–293.

[6] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell, "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health," *Psychiatric Rehabilitation Journal*, vol. 38, no. 3, pp. 218–226, 2015.

[7] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland, "Daily stress recognition from mobile phone data, weather conditions and individual traits," in *Proc. of ACM International Conference on Multimedia*. ACM Press, 2014, pp. 477–486.

[8] A. Bogomolov, B. Lepri, and F. Pianesi, "Happiness recognition from mobile phone data," in *2013 international conference on social computing*. IEEE, 2013, pp. 790–795.

[9] A. Borbely and A. Wirz-Justice, "Sleep, sleep deprivation and depression," *Hum Neurobiol*, vol. 1, no. 205, p. 10, 1982.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] J. T. Cacioppo, L. C. Hawkley, and R. A. Thisted, "Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago health, aging, and social relations study." *Psychology and aging*, vol. 25, no. 2, p. 453, 2010.

[12] Ö. Çağan, A. Ünsal, and N. Çelik, "Evaluation of college students' the level of addiction to cellular phone and investigation on the relationsship between the addiction and the level of depression," *Procedia-Social and Behavioral Sciences*, vol. 114, pp. 831–839, 2014.

[13] L. Canzian and M. Musolesi, "Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proc. of ACM UbiComp*, 2015, pp. 1293–1304.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[16] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive sleep monitoring using smartphones," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 2013, pp. 145–152.

[17] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. of the 36th Annual Symposium on the Foundations of Computer Science (FOCS)*, October 1995, pp. 41–50.

[18] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, 1998.

[19] I. P. Chow, K. Fua, Y. Huang, W. Bonelli, H. Xiong, E. L. Barnes, and A. B. Teachman, "Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students," *J Med Internet Res*, vol. 19, no. 3, p. e62, Mar 2017.

[20] P. Cuijpers and F. Smit, "Excess mortality in depression: A meta-analysis of community studies," *Journal of Affective Disorders*, vol. 72, no. 3, pp. 227–236, 2002.

[21] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *ACM KDD*, 1996.

[22] H. Falaki and S. Keshav, "Trace-based analysis of wi-fi scanning strategies," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 13, no. 1, pp. 73–76, 2009.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[24] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis, "Multi-view bi-clustering to identify smartphone sensing features indicative of depression," in *Proc. IEEE CHASE*, June 2016.

[25] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data," in *Proc. of Wireless Health*, 2016.

[26] M. Faurholt-Jepsen, M. Bauer, and L. V. Kessing, "Smartphone-based objective monitoring in bipolar disorder: status and considerations," *International journal of bipolar disorders*, vol. 6, no. 1, pp. 1–7, 2018.

[27] M. Faurholt-Jepsen, M. Vinberg, M. Frost, S. Debel, E. Margrethe Christensen, J. E. Bardram, and L. V. Kessing, "Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder," *International journal of methods in psychiatric research*, vol. 25, no. 4, pp. 309–323, 2016.

[28] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Supporting disease insight through data analysis: refinements of the monarca self-assessment system," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 133–142.

[29] C. Gentry and Z. Ramzan, "Single-database private information retrieval with constant communication rate," in *Proc. of The 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, vol. 3580, 2005, pp. 803 – 815.

[30] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[31] A. Gruenerbl, V. Osmani, G. Bahle, J. C. Carrasco, S. Oehler, O. Mayora, C. Haring, and P. Lukowicz, "Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients," in *Proceedings of the 5th Augmented Human International Conference*, 2014, pp. 1–8.

[32] A. Grünerbl, P. Oleksy, G. Bahle, C. Haring, J. Weppner, and P. Lukowicz, "Towards smart phone based monitoring of bipolar disorder," in *Proceedings of the Second ACM Workshop on Mobile Systems, Applications, and Services for HealthCare*. ACM, 2012, p. 3.

[33] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[34] G. M. Harari, S. R. Müller, M. S. Aung, and P. J. Rentfrow, "Smartphone sensing methods for studying behavior in everyday life," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 83–90, 2017.

[35] D. Harley, S. Winn, S. Pemberton, and P. Wilcox, "Using texting to support students' transition to university," *Innovations in Education and Teaching International*, vol. 44, no. 3, pp. 229–241, 2007.

[36] H. Hong, C. Luo, and M. C. Chan, "Socialprobe: understanding social interaction through passive wifi monitoring," in *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 2016, pp. 94–103.

[37] W. Hsu, D. Dutta, and A. Helmy, "Extended Abstract: Mining Behavioral Groups in Large Wireless LANs," in *Proc. of ACM MobiCom*, September 2007.

[38] W.-j. Hsu, D. Dutta, and A. Helmy, "Csi: A paradigm for behavior-oriented profile-cast services in mobile networks," *Ad Hoc Networks*, vol. 10, no. 8, pp. 1586–1602, 2012.

[39] S. S. Kanhere, "Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces," in *Mobile Data Management (MDM), 2011 12th IEEE International Conference on*, vol. 2. IEEE, 2011, pp. 3–6.

[40] W. Katon and P. Ciechanowski, "Impact of major depression on chronic medical illness," *J Psychosom Res*, vol. 53, no. 4, pp. 859–863, October 2002.

[41] I. Kawachi and L. F. Berkman, "Social ties and mental health," *Journal of Urban health*, vol. 78, no. 3, pp. 458–467, 2001.

[42] J.-H. Kim, M. Seo, and P. David, "Alleviating depression only to become problematic mobile phone users: Can face-to-face communication be the antidote?" *Computers in Human Behavior*, vol. 51, pp. 440–447, 2015.

[43] K.-H. Kim, A. W. Min, D. Gupta, P. Mohapatra, and J. P. Singh, "Improving energy efficiency of wi-fi sensing on smartphones," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 2930–2938.

[44] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[45] N. Lathia, K. Rachuri, C. Mascolo, and G. Roussos, "Open source smartphone libraries for computational social science," in *Proc. of ACM UbiComp*, ser. UbiComp '13 Adjunct, 2013, pp. 911–920.

[46] S. Lee, C. L. Tam, and Q. T. Chie, "Mobile phone usage preferences: The contributing factors of personality, social anxiety and loneliness," *Social Indicators Research*, vol. 118, no. 3, pp. 1205–1228, 2014.

[47] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong, "Moodscope: Building a mood sensor from smartphone usage patterns," in *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 389–402.

[48] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Advances in neural information processing systems*, 2013, pp. 431–439.

[49] J. Lu, C. Shang, C. Yue, R. Morillo, S. Ware, J. Kamath, A. Bamis, A. Russell, B. Wang, and J. Bi, "Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 21, 2018.

[50] A. Mehrotra, R. Hendley, and M. Musolesi, "Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction," in *Proc. of UbiComp*, 2016.

[51] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2014, pp. 477–486.

[52] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.

[53] ncbi. Social relationships and health: A flashpoint for health policy. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150158/

[54] N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. D. Vos, "Detecting bipolar depression from geographic location data," *IEEE Transactions on Biomedical Engineering*, vol. PP, no. 99, pp. 1–1, 2016.

[55] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[56] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely, "Energy-delay tradeoffs in smartphone applications," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 255–270.

[57] A. Rakotomamonjy, "Variable selection using svm-based criteria," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1357–1370, 2003.

[58] R. Razavi, A. Gharipour, and M. Gharipour, "Depression screening using mobile phone usage metadata: a machine learning approach," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 522–530, 2020.

[59] D. A. Rohani, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review," *JMIR mHealth and uHealth*, vol. 6, no. 8, p. e165, 2018.

[60] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber *et al.*, "The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression," *Biological psychiatry*, vol. 54, no. 5, pp. 573–583, 2003.

[61] S. Saeb, M. Zhang, C. J. Karr, S. M. Schueller, M. E. Corden, K. P. Kording, and D. C. Mohr, "Mobile phone sensor correlates of depressive symptom severity in

daily-life behavior: An exploratory study," *Journal of Medical Internet Research*, vol. 17, no. 7, 2015.

[62] C. E. Sanders, T. M. Field, D. Miguel, and M. Kaplan, "The relationship of Internet use to depression and social isolation among adolescents," *Adolescence*, vol. 35, no. 138, p. 237, 2000.

[63] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on.* IEEE, 2013, pp. 671–676.

[64] C. Segrin, *Interpersonal processes in psychological problems.* Guilford Press, 2001.

[65] S. Servia-Rodríguez, K. K. Rachuri, C. Mascolo, P. J. Rentfrow, N. Lathia, and G. M. Sandstrom, "Mobile sensing at the service of mental well-being: a large-scale longitudinal study," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 103–112.

[66] B. Shumaker and R. Sinnott, "Astronomical computing: 1. computing under the open sky. 2. virtues of the haversine." *Sky and telescope*, vol. 68, pp. 158–159, 1984.

[67] G. Simon, "Social and economic burden of mood disorders," *Biol Psychiatry*, vol. 54, no. 3, pp. 208–215, August 2003.

[68] G. E. Simon, "Social and economic burden of mood disorders," *Biological Psychiatry*, vol. 54, no. 3, pp. 208–215, 2003.

[69] G. E. Simon, C. M. Rutter, D. Peterson, M. Oliver, U. Whiteside, B. Operskalski, and E. J. Ludman, "Does response on the phq-9 depression questionnaire predict subsequent suicide attempt or suicide death?" *Psychiatric Services*, vol. 64, no. 12, pp. 1195–1202, 2013.

[70] K. M. Smith, P. F. Renshaw, and J. Bilello, "The diagnosis of depression: current and emerging methods," *Comprehensive Psychiatry*, vol. 54, no. 1, pp. 1–6, January 2013.

[71] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proc. of WWW*, 2017.

[72] S. Thomée, A. Härenstam, and M. Hagberg, "Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study," *BMC public health*, vol. 11, no. 1, p. 66, 2011.

[73] J. Torous, P. Staples, M. Shanahan, C. Lin, P. Peck, M. Keshavan, and J.-P. Onnela, "Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (phq-9) depressive symptoms in patients with major depressive disorder," *JMIR mental health*, vol. 2, no. 1, 2015.

[74] N. Tsuno, A. Besset, and K. Ritchie, "Sleep and depression." *The Journal of clinical psychiatry*, 2005.

[75] R. J. Turner and W. R. Avison, "Status variations in stress exposure: Implications for the interpretation of research on race, socioeconomic status, and gender," *Journal of Health and Social Behavior*, pp. 488–505, 2003.

[76] D. Umberson and J. Karas Montez, "Social relationships and health: A flashpoint for health policy," *Journal of health and social behavior*, vol. 51, no. 1_suppl, pp. S54–S66, 2010.

[77] A. Višnjić, V. Veličković, D. Sokolović, M. Stanković, K. Mijatović, M. Stojanović, Z. Milošević, and O. Radulović, "Relationship between the manner of mobile phone use and depression, anxiety, and stress in university students," *International journal of environmental research and public health*, vol. 15, no. 4, p. 697, 2018.

[78] T. Vos, A. D. Flaxman, M. Naghavi, R. Lozano, C. Michaud, M. Ezzati, K. Shibuya, J. A. Salomon, S. Abdalla, V. Aboyans *et al.*, "Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010," *The lancet*, vol. 380, no. 9859, pp. 2163–2196, 2012.

[79] R. Wadhwa, A. Chugh, A. Kumar, M. Singh, K. Yadav, S. Eswaran, and T. Mukherjee, "Sensex: Design and deployment of a pervasive wellness monitoring platform for workplaces," in *International Conference on Service-Oriented Computing*. Springer, 2015, pp. 427–443.

[80] R. Wang, M. S. H. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhuryy, M. Hauserz, J. Kanez, M. Merrilly, E. A. Scherer, V. W. S. Tsengy, and D. Ben-Zeev, "Crosscheck: Toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proc. of UbiComp*, 2016.

[81] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing mental health, academic

performance and behavioral trends of college students using smartphones," in *Proc. of ACM Ubicomp*, 2014, pp. 3–14.

[82] R. Wang, W. Wang, A. daSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[83] S. Ware, C. Yue, R. Morillo, J. Lu, C. Shang, J. Kamath, A. Bamis, J. Bi, A. Russell, and B. Wang, "Large-scale automatic depression screening using meta-data from wifi infrastructure," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–27, 2018.

[84] A. Wirz-Justice and R. H. Van den Hoofdakker, "Sleep deprivation in depression: what do we know, where do we go?" *Biological psychiatry*, vol. 46, no. 4, pp. 445–453, 1999.

[85] X. Xu, P. Chikersal, A. Doryab, D. K. Villalba, J. M. Dutcher, M. J. Tumminia, T. Althoff, S. Cohen, K. G. Creswell, J. D. Creswell *et al.*, "Leveraging routine behavior and contextually-filtered features for depression detection among college students," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 3, pp. 1–33, 2019.

[86] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.

[87] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, A. Russell, A. Bamis, and B. Wang, "Fusing location data for depression prediction," in *Proc. IEEE Ubiquitous Intelligence and Computing*, August 2017.

[88] ——, "Fusing location data for depression prediction," *IEEE Transactions on Big Data*, October 2018.

[89] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Fusing location data for depression prediction," *IEEE Transactions on Big Data*, 2018.

[90] H. Zhang, Z. Yan, J. Yang, E. M. Tapia, and D. J. Crandall, "Mfingerprint: Privacy-preserving user modeling with multimodal mobile device footprints," in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*. Springer, 2014, pp. 195–203.

[91] D. Zhou, J. Luo, V. M. B. Silenzio, Y. Zhou, J. Hu, G. Currier, and H. A. Kautz, "Tackling mental health by integrating unobtrusive multimodal sensing," in *Proc. of AAAI*, 2015.